RESEARCH CENTER

FIELD
Perception, Cognition and Interaction

Activity Report 2019

Section Scientific Foundations

Edition: 2020-03-21

CEDAR Project-Team

3. Research Program

3.1. Scalable Heterogeneous Stores

Big Data applications increasingly involve *diverse* data sources, such as: structured or unstructured documents, data graphs, relational databases etc. and it is often impractical to load (consolidate) diverse data sources in a single repository. Instead, interesting data sources need to be exploited "as they are", with the added value of the data being realized especially through the ability to combine (join) together data from several sources. Systems capable of exploiting diverse Big Data in this fashion are usually termed *polystores*. A current limitation of polystores is that data stays captive of its original storage system, which may limit the data exploitation performance. We work to devise highly efficient storage systems for heterogeneous data across a variety of data stores.

3.2. Semantic Query Answering

In the presence of data semantics, query evaluation techniques are insufficient as they only take into account the database, but do not provide the reasoning capabilities required in order to reflect the semantic knowledge. In contrast, (ontology-based) query answering takes into account both the data and the semantic knowledge in order to compute the full query answers, blending query evaluation and semantic reasoning.

We aim at designing efficient semantic query answering algorithms, both building on cost-based reformulation algorithms developed in the team and exploring new approaches mixing materialization and reformulation.

3.3. Multi-Model Querying

As the world's affairs get increasingly more digital, a large and varied set of data sources becomes available: they are either structured databases, such as government-gathered data (demographics, economics, taxes, elections, ...), legal records, stock quotes for specific companies, un-structured or semi-structured, including in particular graph data, sometimes endowed with semantics (see e.g. the Linked Open Data cloud). Modern data management applications, such as data journalism, are eager to combine in innovative ways both static and dynamic information coming from structured, semi-structured, and un-structured databases and social feeds. However, current content management tools for this task are not suited for the task, in particular when they require a lenghy rigid cycle of data integration and consolidation in a warehouse. Thus, we see a need for flexible tools allowing to interconnect various kinds of data sources and to query them together.

3.4. Interactive Data Exploration at Scale

In the Big Data era we are faced with an increasing gap between the fast growth of data and the limited human ability to comprehend data. Consequently, there has been a growing demand of data management tools that can bridge this gap and help users retrieve high-value content from data more effectively. To respond to such user information needs, we aim to build interactive data exploration as a new database service, using an approach called "explore-by-example".

3.5. Exploratory Querying of Semantic Graphs

Semantic graphs including data and knowledge are hard to apprehend for users, due to the complexity of their structure and oftentimes to their large volumes. To help tame this complexity, in prior research (2014), we have presented a full framework for RDF data warehousing, specifically designed for heterogeneous and semantic-rich graphs. However, this framework still leaves to the users the burden of chosing the most interesting warehousing queries to ask. More user-friendly data management tools are needed, which help the user discover the interesting structure and information hidden within RDF graphs. This research has benefitted from the arrival in the team of Mirjana Mazuran, as well as from the start of the PhD thesis of Paweł Guzewicz, co-advised by Yanlei Diao and Ioana Manolescu.

3.6. An Unified Framework for Optimizing Data Analytics

Data analytics in the cloud has become an integral part of enterprise businesses. Big data analytics systems, however, still lack the ability to take user performance goals and budgetary constraints for a task, collectively referred to as task objectives, and automatically configure an analytic job to achieve the objectives.

Our goal, is to come up with a data analytics optimizer that can automatically determine a cluster configuration with a suitable number of cores as well as other runtime system parameters that best meet the task objectives. To achieve this, we also need to design a multi-objective optimizer that constructs a Pareto optimal set of job configurations for task-specific objectives, and recommends new job configurations to best meet these objectives.

GRAPHIK Project-Team

3. Research Program

3.1. Logic-based Knowledge Representation and Reasoning

We follow the mainstream *logic-based* approach to knowledge representation (KR). First-order logic (FOL) is the reference logic in KR and most formalisms in this area can be translated into fragments (i.e., particular subsets) of FOL. This is in particular the case for description logics and existential rules, two well-known KR formalisms studied in the team.

A large part of research in this domain can be seen as studying the *trade-off* between the expressivity of languages and the complexity of (sound and complete) reasoning in these languages. The fundamental problem in KR languages is entailment checking: is a given piece of knowledge entailed by other pieces of knowledge, for instance from a knowledge base (KB)? Another important problem is *consistency* checking: is a set of knowledge pieces (for instance the knowledge base itself) consistent, i.e., is it sure that nothing absurd can be entailed from it? The *ontology-mediated query answering* problem is a topical problem (see Section 3.3). It asks for the set of answers to a query in the KB. In the case of Boolean queries (i.e., queries with a yes/no answer), it can be recast as entailment checking.

3.2. Graph-based Knowledge Representation and Reasoning

Besides logical foundations, we are interested in KR formalisms that comply, or aim at complying with the following requirements: to have good *computational* properties and to allow users of knowledge-based systems to have a maximal *understanding and control* over each step of the knowledge base building process and use.

These two requirements are the core motivations for our graph-based approach to KR. We view labelled graphs as an *abstract representation* of knowledge that can be expressed in many KR languages (different kinds of conceptual graphs —historically our main focus— the Semantic Web language RDF (Resource Description Framework), its extension RDFS (RDF Schema), expressive rules equivalent to the so-called tuple-generating-dependencies in databases, some description logics dedicated to query answering, etc.). For these languages, reasoning can be based on the structure of objects, thus based on graph-theoretic notions, while staying logically founded.

More precisely, our basic objects are labelled graphs (or hypergraphs) representing entities and relationships between these entities. These graphs have a natural translation in first-order logic. Our basic reasoning tool is graph homomorphism. The fundamental property is that graph homomorphism is sound and complete with respect to logical entailment *i.e.*, given two (labelled) graphs G and H, there is a homomorphism from G to H *if and only if* the formula assigned to G is entailed by the formula assigned to H. In other words, logical reasoning on these graphs can be performed by graph mechanisms. These knowledge constructs and the associated reasoning mechanisms can be extended (to represent rules for instance) while keeping this fundamental correspondence between graphs and logics.

3.3. Ontology-Mediated Query Answering

Querying knowledge bases has become a central problem in knowledge representation and in databases. A knowledge base (KB) is classically composed of a terminological part (metadata, ontology) and an assertional part (facts, data). Queries are supposed to be at least as expressive as the basic queries in databases, i.e., conjunctive queries, which can be seen as existentially closed conjunctions of atoms or as labelled graphs. The challenge is to define good trade-offs between the expressivity of the ontological language and the complexity of querying data in presence of ontological knowledge. Description logics have been so far the prominent family of formalisms for representing and reasoning with ontological knowledge. However, classical description logics were not designed for efficient data querying. On the other hand, database languages are able to process complex queries on huge databases, but without taking the ontology into account. There is thus a need for new languages and mechanisms, able to cope with the ever growing size of knowledge bases in the Semantic Web or in scientific domains.

This problem is related to two other problems identified as fundamental in KR:

- *Query answering with incomplete information.* Incomplete information means that it might be unknown whether a given assertion is true or false. Databases classically make the so-called closed-world assumption: every fact that cannot be retrieved or inferred from the base is assumed to be false. Knowledge bases classically make the open-world assumption: if something cannot be inferred from the base, and neither can its negation, then its truth status is unknown. The need of coping with incomplete information is a distinctive feature of querying knowledge bases with respect to querying classical databases (however, as explained above, this distinction tends to disappear). The presence of incomplete information makes the query answering task much more difficult.
- *Reasoning with rules.* Researching types of rules and adequate manners to process them is a mainstream topic in the Semantic Web, and, more generally a crucial issue for knowledge-based systems. For several years, we have been studying rules, both in their logical and their graph form, which are syntactically very simple but also very expressive. These rules, known as existential rules or Datalog+, can be seen as an abstraction of ontological knowledge expressed in the main languages used in the context of KB querying.

3.4. Inconsistency and Decision Making

While classical FOL is the kernel of many KR languages, to solve real-world problems we often need to consider features that cannot be expressed purely (or not naturally) in classical logic. The logic and graph-based formalisms used for previous points have thus to be extended with such features. The following requirements have been identified from scenarios in decision making, privileging the agronomy domain:

- to cope with inconsistency;
- to cope with defeasible knowledge;
- to take into account different and potentially conflicting viewpoints;
- to integrate decision notions (priorities, gravity, risk, benefit).

Although the solutions we develop require to be validated on the applications that motivated them, we also want them to be sufficiently generic to be applied in other contexts. One angle of attack (but not the only possible one) consists in increasing the expressivity of our core languages, while trying to preserve their essential combinatorial properties, so that algorithmic optimizations can be transferred to these extensions.

LACODAM Project-Team

3. Research Program

3.1. Introduction

The three original research axes of the LACODAM project-team are the following. First, we briefly introduce these axes, as well as their interplay. We then introduce the axis of *Interpretable AI* (Section 3.4), whose emergence is a response to the current societal needs.

- The first research axis (Section 3.2) is dedicated to the design of *novel pattern mining methods*. Pattern mining is one of the most important approaches to discover novel knowledge in data, and one of our strongest areas of expertise. The work on this axis will serve as foundations for work on the other two axes. Thus, this axis will have the strongest impact on our overall goals.
- The second axis (Section 3.3) tackles another aspect of knowledge discovery in data: the *interaction between the user and the system* in order to co-discover novel knowledge. Our team has plenty of experience collaborating with domain experts, and is therefore aware of the need to improve such interaction.
- The third axis (Section 3.4) concerns *decision support*. With the help of methods from the two previous axes, our goal here is to design systems that can either assist humans with making decisions, or make relevant decisions in situations where extremely fast reaction is required.

Figure 1 sums up the detailed work presented in the next few pages: we show the three research axes of the team (X-axis) on the left and our main applications areas (Y-axis) below. In the middle there are colored squares that represent the precise research topics of the team aligned with their axis and main application area. These research topics will be described in this section. Lines represent projects that can link several topics, and that are also connected to their main application area.

3.2. Pattern mining algorithms

Twenty years of research in pattern mining have resulted in efficient approaches to handle the algorithmic complexity of the problem. Existing algorithms are now able to efficiently extract patterns with complex structures (ex: sequences, graphs, co-variations) from large datasets. However, when dealing with large, real-world datasets, these methods still output a huge set of patterns, which is impractical for human analysis. This problem is called *pattern explosion*. The ongoing challenge of pattern mining research is to extract fewer but more meaningful patterns. The LACODAM team is committed to solve the pattern explosion problem by pursuing the following four research topics:

1. the design of dedicated algorithms for mining temporal patterns
2. the design of flexible pattern mining approaches
3. the automatic selection of interesting data mining results
4. the design of parallel pattern algorithms to ensure scalability

The originality of our contributions relies on the exploration of knowledge-based approaches whose principle is to incorporate dedicated domain knowledge (aka application background knowledge) deep into the mining process. While most data mining approaches are based on agnostic approaches designed to cope with pattern explosion, we propose to develop data mining techniques that rely on knowledge-based artificial intelligence techniques. This entails the use of structured knowledge representations, as well as reasoning methods, in combination with mining.

The first topic concerns classical pattern mining in conjunction with expert knowledge in order to define new pattern types (and related algorithms) that can solve applicative issues. In particular, we investigate how to handle temporality in pattern representations which turns out to be important in many real world applications (in particular for decision support) and deserves particular attention.

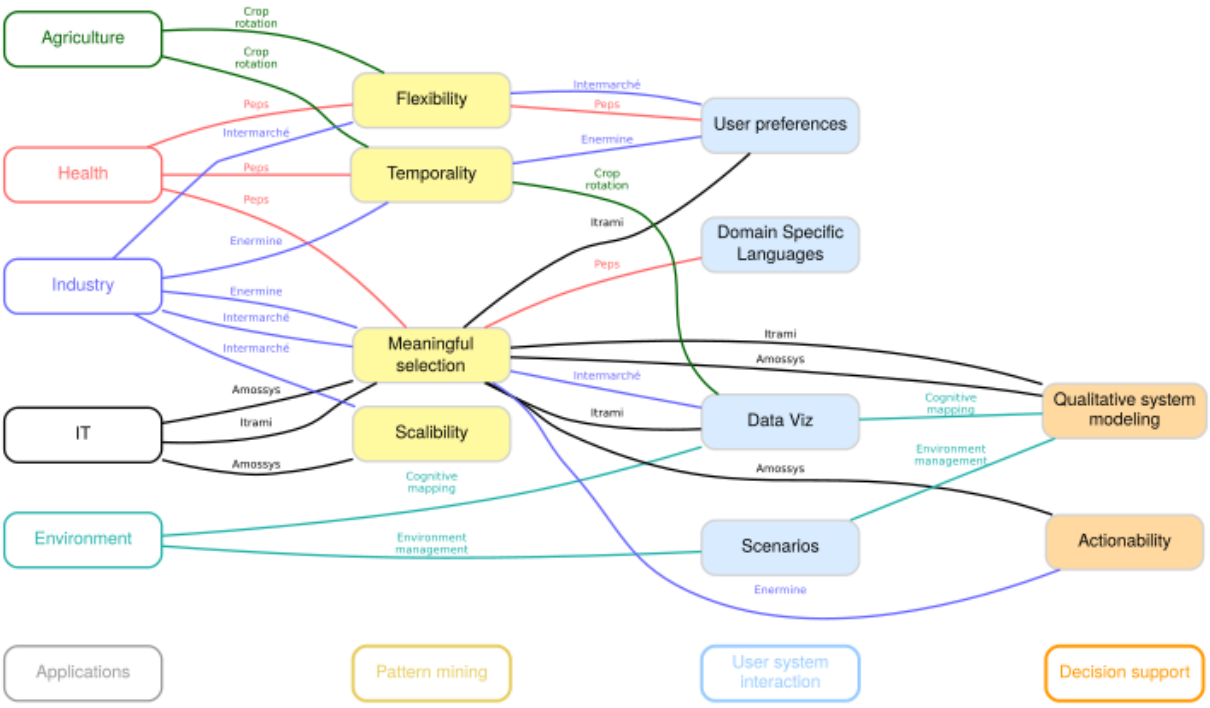


Figure 1. LACODAM research topics organized by axis and application

The next two topics aim at proposing alternative pattern mining methods to let the user incorporate, on her own, knowledge that will help define her pattern domain of interest. Flexible pattern mining approaches enable analysts to easily incorporate extra knowledge, for example domain related constraints, in order to extract only the most relevant patterns. On the other hand, the selection of interesting data mining results aims at devising strategies to filter out the results that are useless to the data analyst. Besides the challenge of algorithmic efficiency, we are interested in formalizing the foundations of interestingness, according to background knowledge modeled with logical knowledge representation paradigms.

Last but not least, pattern mining algorithms are compute-intensive. It is thus important to exploit all the available computing power. Parallelism is for a foreseeable future one of the main ways to speed up computations, and we have a strong competence on the design of parallel pattern mining algorithms. We will exploit this competence in order to guarantee that our approaches scale up to the data provided by our partners.

3.3. User/system interaction

As we pointed out before, there is a strong need to present relevant patterns to the user. This can be done by using more specific constraints, background knowledge and/or tailor-made optimization functions. Due to the difficulty of determining these elements beforehand, one of the most promising solutions is that the system and the user co-construct the definition of relevance, i.e., to have a human in the loop. This requires to have means to present intermediate results to the user, and to get user feedback in order to guide the search space exploration process in the right direction. This is an important research axis for LACODAM, which will be tackled in several complementary ways:

- *Domain Specific Languages:* One way to interact with the user is to propose a Domain Specific Language (DSL) tailored to the domain at hand and to the analysis tasks. The challenge is to propose a DSL allowing the users to easily express the required processing workflows, to deploy those workflows for mining on large volumes of data and to offer as much automation as possible.
- *What if / What for scenarios:* We also investigate the use of scenarios to query results from data mining processes, as well as other complex processes such as complex system simulations or model predictions. Such scenarios are answers to questions of the type “what if [situation]?” or “what [should be done] for [expected outcome]?”.
- *User preferences:* In exploratory analysis, users often do not have a precise idea of what they want, and are not able to formulate such queries. Hence, in LACODAM we investigate simple ways for users to express their interests and preferences, either during the mining process – to guide the search space exploration –, or afterwards during the filtering and interpretation of the most relevant results.
- *Data visualization:* Most of the research directions presented in this document require users to examine patterns at some point. The output of most pattern mining algorithms is usually a (long) list of patterns. While this presentation can be sufficient for some applications, often it does not provide a complete understanding, especially for non-experts in pattern mining. A transversal research topic that we want to explore in LACODAM is to propose data visualization techniques that are adequate for understanding output results. Numerous (failed) experiments have shown that data mining and data visualization are fields, which require distinct skills, thus researchers in one field usually do not make significant advances in the other field (this is detailed in [Keim 2010]). Thus, our strategy is to establish collaborations with prominent data visualization teams for this line of research, with a long term goal to recruit a specialist in data visualization if the opportunity arises.

3.4. Decision support

Patterns have proved to be quite useful for decision-aid. Predictive sequential patterns, to give an example, have a direct application in diagnosis. Itemsets and contrast patterns can be used for interpretable machine learning (ML). In regards to diagnosis, LACODAM inherits, from the former DREAM team, a strong background in decision support systems with internationally recognized expertise in this field. This subfield of AI (Artificial Intelligence) is concerned with determining whether a system is operating normally or not, and the cause of

faulty behaviors. The studied system can be an agro- or eco-system, a software system (e.g., a ML classifier), a living being, etc. In relation to interpretable machine learning (ML), this subfield is concerned with the conception of models whose answers are understandable by users. This can be achieved by inducing inherently white-box models from data such as rule-based classifiers/regressors, or by mining rules and explanations from black-box models. The latter setting is quite common due to the high accuracy of black-box models compared to natively interpretable models. Pattern mining is a powerful tool to mine explanations from black-box systems. Those explanations can be used to diagnose biases in systems, either to debug and improve the model, or to generate trust in the verdicts of intelligent software agents.

The increasing volumes of data coming from a range of different systems (ex: sensor data from agro-environmental systems, log data from software systems and ML models, biological data coming from health monitoring systems) can help human and software agents make better decisions. Hence, LACODAM builds upon the idea that decision support systems (an interest bequeathed from DREAM) should take advantage of the available data. This third and last research axis is thus a meeting point for all members of the team, as it requires the integration of AI techniques for traditional decision support systems with results from data mining techniques.

Three main research sub-axes are investigated in LACODAM:

- *Diagnosis-based approaches.* We are exploring how to integrate knowledge found from pattern mining approaches, possibly with the help of interactive methods, into the qualitative models. The goal of such work is to automate as much as possible the construction of prediction models, which can require a lot of human effort.
- *Actionable patterns and rules.* In many settings of "exploratory data mining", the actual interestingness of a pattern is hard to assess, as it may be subjective. However, for some applications there are well defined measures of interestingness and applicability for patterns. Patterns and rules that can lead to actual actions –that are relevant to the user– are called "actionable patterns" and are of vital importance to industrial settings.
- *Mining explanations from ML systems.* Interpretable ML and AI is a current trend for technical, ethical, and legal reasons [27]. In this regard, pattern mining can be used to spot regularities that arise when a complex black-box model yields a particular verdict. For instance, one may want to know the conditions under which the control module of a self-driving car decided to stop without apparent reason, or which factors caused a ML-based credit assessor to reject a loan request. Patterns and conditions are the building blocks for the generation of human-readable explanations for such black-box systems.

3.5. Interpretability

The pervasiveness of complex decision support systems, as well as the general consensus about the societal importance of understanding the rationale embedded in such systems [0], has given momentum to the field of interpretable ML. Being a team specialized in data science, we are fully aware that many problems can be solved by means of complex and accurate ML models. Alas, this accuracy sometimes comes at the expense of interpretability, which can be a major requirement in some contexts (e.g., regression using expertise/rule mining). For this reason, one of the interests of LACODAM is the study of the interpretability-accuracy trade-off. Our studies may be able to answer questions such as "how much accuracy can a model lose (or perhaps gain) by becoming more interpretable?". Such a goal requires us to define interpretability in a more principled way —an endeavour that has been very recently addressed, still not solved. LACODAM is interested in the two main currents of research in interpretability, namely the development of natively interpretable methods, as well as the construction of interpretable mediators between users and black-box models, known as post-hoc interpretability.

[0] General Data Protection Regulation, recital 71 http://www.privacy-regulation.eu/en/r71.htm

We highlight the link between interpretability and LACODAM's axes of decision support, and user/system interaction. In particular, interpretability is a prerequisite for proper user/system interaction and is a central incentive for the advent of data visualization techniques for ML models. This convergence has motivated our interest in *user-oriented post-hoc interpretability*, a sub-field of interpretable ML that adds the user into the formula when generating proper explanations of black-box ML algorithms. This rationale is supported by existing work [28] that suggests that interpretability possesses a subjective component known as plausibility. Moreover, our user-oriented vision meets with the notion of semantic interpretability, where an explanation may resort to high level semantic elements (objects in image classification, or verbal phases in natural language processing) instead of surrogate still-machine-friendly features (such as super-pixels). LACODAM will tackle all these unaddressed aspects of interpretable ML with other Inria teams through the IPL HyAIAI.

3.6. Long-term goals

The following perspectives are at the convergence of the four aforementioned research axes and can be seen as ideal towards our goals:

- *Automating data science workflow discovery.* The current methods for knowledge extraction and construction of decision support systems require a lot of human effort. Our three research axes aim at alleviating this effort, by devising methods that are more generic and by improving the interaction between the user and the system. An ideal solution would be that the user could forget completely about the existence of pattern mining or decision support methods. Instead the user would only loosely specify her problem, while the system constructs various data science / decision support workflows, possibly further refined via interactions.

  We consider that this is a second order AI task, where AI techniques such as planning are used to explore the workflow search space, the workflow itself being composed of data mining and/or decision support components. This is a strategic evolution for data science endeavors, were the demand far exceeds the available human skilled manpower.

- *Logic argumentation based on epistemic interest.* Having increasingly automated approaches will require better and better ways to handle the interactions with the user. Our second long term goal is to explore the use of logic argumentation, i.e., the formalisation of human strategies for reasoning and arguing, in the interaction between users and data analysis tools. Alongside visualization and interactive data mining tools, logic argumentation can be a way for users to query both the results and the way they are obtained. Such querying can also help the expert to reformulate her query in an interactive analysis setting.

  This research direction aims at exploiting principles of interactive data analysis in the context of epistemic interestingness measures. Logic argumentation can be a natural tool for interactions between the user and the system: display of possibly exhaustive list of arguments, relationships between arguments (e.g., reinforcement, compatibility or conflict), possible solutions for argument conflicts, etc.

  The first step is to define a formal argumentation framework for explaining data mining results. This implies to continue theoretical work on the foundations of argumentation in order to identify the most adapted framework (either existing or a new one to be defined). Logic argumentation may be implemented and deeply explored in ASP, allowing us to build on our expertise in this logic language.

- *Collaborative feedback and knowledge management.* We are convinced that improving the data science process, and possibly automating it, will rely on high-quality feedback from communities on the web. Consider for example what has been achieved by collaborative platforms such as StackOverflow: it has become the reference site for any programming question.

  Data science is a more complex problem than programming, as in order to get help from the community, the user has to share her data and workflow, or at least some parts of them. This raises obvious privacy issues that may prevent this idea to succeed. As our research on automating the

production of data science workflows should enable more people to have access to data science results, we are interested in the design of collaborative platforms to exchange expert advices over data, workflows and analysis results. This aims at exploiting human feedback to improve the automation of data science system via machine learning methods.

LINKS Project-Team

3. Research Program

3.1. Background

The main objective of LINKS is to develop methods for querying and managing linked data collections. Even though open linked data is the most prominent example, we will focus on hybrid linked data collections, which are collections of semi-structured datasets in hybrid formats: graph-based, RDF, relational, and NOSQL. The elements of these datasets may be linked, either by pointers or by additional relations between the elements of the different datasets, for instance the "same-as" or "member-of" relations as in RDF.

The advantage of traditional data models is that there exist powerful querying methods and technologies that one might want to preserve. In particular, they come with powerful schemas that constraint the possible manners in which knowledge is represented to a finite number of patterns. The exhaustiveness of these patterns is essential for writing of queries that cover all possible cases. Pattern violations are excluded by schema validation. In contrast, RDF schema languages such as RDFS can only enrich the relations of a dataset by new relations, which also helps for query writing, but which cannot constraint the number of possible patterns, so that they do not come with any reasonable notion of schema validation.

The main weakness of traditional formats, however, is that they do not scale to large data collections as stored on the Web, while the RDF data models scales well to very big collections such as linked open data. Therefore, our objective is to study mixed data collections, some of which may be in RDF format, in which we can lift the advantages of smaller datasets in traditional formats to much larger linked data collections. Such data collections are typically distributed over the internet, where data sources may have rigid query facilities that cannot be easily adapted or extended.

The main assumption that we impose in order to enable the logical approach, is that the given linked data collection must be correct in most dimensions. This means that all datasets are well-formed with respect to their available constraints and schemas, and clean with respect to the data values in most of the components of the relations in the datasets. One of the challenges is to integrate good quality RDF datasets into this setting, another is to clean the incorrect data in those dimensions that are less proper. It remains to be investigated in how far these assumptions can be maintained in realistic applications, and how much they can be weakened otherwise.

For querying linked data collections, the main problems are to resolve the heterogeneity of data formats and schemas, to understand the efficiency and expressiveness of recursive queries, that can follow links repeatedly, to answer queries under constraints, and to optimize query answering algorithms based on static analysis. When linked data is dynamically created, exchanged, or updated, the problems are how to process linked data incrementally, and how to manage linked data collections that change dynamically. In any case (static and dynamic) one needs to find appropriate schema mappings for linking semi-structured datasets. We will study how to automatize parts of this search process by developing symbolic machine learning techniques for linked data collections.

3.2. Querying Heterogeneous Linked Data

Our main objective is to query collections of linked datasets. In the static setting, we consider two kinds of links: explicit links between elements of the datasets, such as equalities or pointers, and logical links between relations of different datasets such as schema mappings. In the dynamic setting, we permit a third kind of links that point to "intentional" relations computable from a description, such as the application of a Web service or the application of a schema mapping.

We believe that collections of linked datasets are usually too big to ensure a global knowledge of all datasets. Therefore, schema mappings and constraints should remain between pairs of datasets. Our main goal is to be able to pose a query on a collection of datasets, while accounting for the possible recursive effects of schema mappings. For illustration, consider a ring of datasets D_1, D_2, D_3 linked by schema mappings M_1, M_2, M_3 that tell us how to complete a database D_i by new elements from the next database in the cycle.

The mappings M_i induce three intentional datasets I_1, I_2, and I_3, such that I_i contains all elements from D_i and all elements implied by M_i from the next intentional dataset in the ring:

$$I_1 = D_1 \cup M_1(I_2), \quad I_2 = D_2 \cup M_2(I_3), \quad I_3 = D_3 \cup M_3(I_1)$$

Clearly, the global information collected by the intentional datasets depends recursively on all three original datasets D_i. Queries to the global information can now be specified as standard queries to the intentional databases I_i. However, we will never materialize the intentional databases I_i. Instead, we can rewrite queries on one of the intentional datasets I_i to recursive queries on the union of the original datasets D_1, D_2, and D_3 with their links and relations. Therefore, a query answering algorithm is needed for recursive queries, that chases the "links" between the D_i in order to compute the part of I_i needed for the purpose of query answering.

This illustrates that we must account for the graph data models when dealing with linked data collections whose elements are linked, and that query languages for such graphs must provide recursion in order to chase links. Therefore, we will have to study graph databases with recursive queries, such as RDF graphs with SPARQL queries, but also other classes of graph databases and queries.

We study schemas and mappings between datasets with different kinds of data models and the complexity of evaluating recursive queries over graphs. In order to use schema mapping for efficiently querying the different datasets, we need to optimize the queries by taking into account the mappings. Therefore, we will study static analysis of schema mappings and recursive queries. Finally, we develop concrete applications in which our fundamental techniques can be applied.

3.3. Managing Dynamic Linked Data

With the quick growth of the information technology on the Web, more and more Web data gets created dynamically every day, for instance by smartphones, industrial machines, users of social networks, and all kinds of sensors. Therefore, large amounts of dynamic data need to be exchanged and managed by various data-centric web services, such as online shops, online newspapers, and social networks.

Dynamic data is often created by the application of some kind of service on the Web. This kind of data is intentional in the same spirit as the intentional data specified by the application of a schema mapping, or the application of some query to the hidden Web. Therefore, we will consider a third kind of links in the dynamic setting, that map to intentional data specified by whatever kind of function application. Such a function can be defined in data-centric programming languages, in the style of Active XML, XSLT, and NoSQL languages.

The dynamicity of data adds a further dimension to the challenges for linked data collections that we described before, while all the difficulties remain valid. One of the new aspects is that intentional data may be produced incrementally, as for instance when exchanged over data streams. Therefore, one needs incremental algorithms able to evaluate queries on incomplete linked data collections, that are extended or updated incrementally. Note that incremental data may be produced without end, such as a Twitter stream, so that one cannot wait for its completion. Instead, one needs to query and manage dynamic data with as low latency as possible. Furthermore, all static analysis problems are to be re-investigated in the presence of dynamic data.

Another aspect of dynamic data is distribution over the Web, and thus parallel processing as in the cloud. This raises the typical problems coming with data distribution: huge data sources cannot be moved without very high costs, while data must be replicated for providing efficient parallel access. This makes it difficult, if not impossible, to update replicated data consistently. Therefore, the consistency assumption has been removed by NoSQL databases for instance, while parallel algorithmic is limited to naive parallelization (i.e. map/reduce) where only few data needs to be exchanged.

We will investigate incremental query evaluation for distributed data-centered programming languages for linked data collections, dynamic updates as needed for linked data management, and static analysis for linked data workflows.

3.4. Linking Graphs

When datasets from independent sources are not linked with existing schema mappings, we would like to investigate symbolic machine learning solutions for inferring such mappings in order to define meaningful links between data from separate sources. This problem can be studied for various kinds of linked data collections. Before presenting the precise objectives, we will illustrate our approach on the example of linking data in two independent graphs: an address book of a research institute containing detailed personnel information and a (global) bibliographic database containing information on papers and their authors.

We remind that a schema allows to identify a collection of types each grouping objects from the same semantic class e.g., the collection of all persons in the address book and the collection of all authors in the bibliography database. As a schema is often lacking or underspecified in graph data models, we intend to investigate inference methods based on structural similarity of graph fragments used to describe objects from the same class in a given document e.g., in the bibliographic database every author has a name and a number of affiliations, while a paper has a title and a number of authors. Furthermore, our inference methods will attempt to identify, for every type, a set of possible keys, where by key we understand a collection of attributes of an object that uniquely identifies such an object in its semantic class. For instance, for a person in the address book two examples of a key are the name of the person and the office phone number of that person.

In the next step, we plan to investigate employing existing entity linkage solutions to identify pairs of types from different databases whose instances should be linked using compatible keys. For instance, persons in the address book should be linked with authors in the bibliographical database using the name as the compatible key. Linking the same objects (represented in different ways) in two databases can be viewed as an instance of a mapping between the two databases. Such mapping is, however, discriminatory because it typically maps objects from a specific subset of objects of given types. For instance, the mapping implied by linking persons in the address book with authors in the bibliographic database involves in fact researchers, a subgroup of personnel of the research institute, and authors affiliated with the research institute. Naturally, a subset of objects of a given type, or a subtype, can be viewed as a result of a query on the set of all objects, which on very basic level illustrates how learning data mappings can be reduced to learning queries.

While basic mappings link objects of the same type, more general mappings define how the same type of information is represented in two different databases. For instance, the email address and the postal address of an individual may be represented in one way in the address book and in another way in the bibliographic databases, and naturally, the query asking for the email address and the postal address of a person identified by a given name will differ from one database to the other. While queries used in the context of linking objects of compatible types are essentially unary, queries used in the context of linking information are n-ary and we plan to approach inference of general database mappings by investigating and employing algorithms for inference of n-ary queries.

An important goal in this research is elaborating a formal definition of *learnability* (feasibility of inference) of a given class of concepts (schemas of queries). We plan to following the example of Gold (1967), which requires not only the existence of an efficient algorithm that infers concepts consistent with the given input but the ability to infer every concept from the given class with a sufficiently informative input. Naturally, learnability depends on two parameters. The first parameter is the class of concepts i.e., a class of schema and

a class of queries, from which the goal concept is to be inferred. The second parameter is the type of input that an inference algorithm is given. This can be a set of examples of a concept e.g., instances of RDF databases for which we wish to construct a schema or a selection of nodes that a goal query is to select. Alternatively, a more general interactive scenario can be used where the learning algorithm inquires the user about the goal concept e.g., by asking to indicate whether a given node is to be selected or not (as membership queries of Angluin (1987)). In general, the richer the input is, the richer class of concepts can be handled, however, the richer class of queries is to be handled, the higher computational cost is to be expected. The primary task is to find a good compromise and identify classes of concepts that are of high practical value, allow efficient inference with possibly simple type of input.

The main open problem for graph-shaped data studied by Links are how to infer queries, schemas, and schema-mappings for graph-structured data.

MAGNET Project-Team

3. Research Program

3.1. Introduction

The main objective of MAGNET is to develop original machine learning methods for networked data in order to build applications like browsing, monitoring and recommender systems, and more broadly information extraction in information networks. We consider information networks in which the data consist of both feature vectors and texts. We model such networks as (multiple) (hyper)graphs wherein nodes correspond to entities (documents, spans of text, users, ...) and edges correspond to relations between entities (similarity, answer, co-authoring, friendship, ...). Our main research goal is to propose new online and batch learning algorithms for various problems (node classification / clustering, link classification / prediction) which exploit the relationships between data entities and, overall, the graph topology. We are also interested in searching for the best hidden graph structure to be generated for solving a given learning task. Our research will be based on generative models for graphs, on machine learning for graphs and on machine learning for texts. The challenges are the dimensionality of the input space, possibly the dimensionality of the output space, the high level of dependencies between the data, the inherent ambiguity of textual data and the limited amount of human labeling. An additional challenge will be to design scalable methods for large information networks. Hence, we will explore how sampling, randomization and active learning can be leveraged to improve the scalability of the proposed algorithms.

Our research program is organized according to the following questions:

1. How to go beyond vectorial classification models in Natural Language Processing (NLP) tasks?
2. How to adaptively build graphs with respect to the given tasks? How to create networks from observations of information diffusion processes?
3. How to design methods able to achieve a good trade-off between predictive accuracy and computational complexity?
4. How to go beyond strict node homophilic/similarity assumptions in graph-based learning methods?

3.2. Beyond Vectorial Models for NLP

One of our overall research objectives is to derive graph-based machine learning algorithms for natural language and text information extraction tasks. This section discusses the motivations behind the use of graph-based ML approaches for these tasks, the main challenges associated with it, as well as some concrete projects. Some of the challenges go beyond NLP problems and will be further developed in the next sections. An interesting aspect of the project is that we anticipate some important cross-fertilizations between NLP and ML graph-based techniques, with NLP not only benefiting from but also pushing ML graph-based approaches into new directions.

Motivations for resorting to graph-based algorithms for texts are at least threefold. First, online texts are organized in networks. With the advent of the web, and the development of forums, blogs, and micro-blogging, and other forms of social media, text productions have become strongly connected. Interestingly, NLP research has been rather slow in coming to terms with this situation, and most of the literature still focus on document-based or sentence-based predictions (wherein inter-document or inter-sentence structure is not exploited). Furthermore, several multi-document tasks exist in NLP (such as multi-document summarization and cross-document coreference resolution), but most existing work typically ignore document boundaries and simply apply a document-based approach, therefore failing to take advantage of the multi-document dimension [38], [41].

A second motivation comes from the fact that most (if not all) NLP problems can be naturally conceived as graph problems. Thus, NLP tasks often involve discovering a relational structure over a set of text spans (words, phrases, clauses, sentences, etc.). Furthermore, the *input* of numerous NLP tasks is also a graph; indeed, most end-to-end NLP systems are conceived as pipelines wherein the output of one processor is in the input of the next. For instance, several tasks take POS tagged sequences or dependency trees as input. But this structured input is often converted to a vectorial form, which inevitably involves a loss of information.

Finally, graph-based representations and learning methods appear to address some core problems faced by NLP, such as the fact that textual data are typically not independent and identically distributed, they often live on a manifold, they involve very high dimensionality, and their annotations is costly and scarce. As such, graph-based methods represent an interesting alternative to, or at least complement, structured prediction methods (such as CRFs or structured SVMs) commonly used within NLP. Graph-based methods, like label propagation, have also been shown to be very effective in semi-supervised settings, and have already given some positive results on a few NLP tasks [21], [43].

Given the above motivations, our first line of research will be to investigate how one can leverage an underlying network structure (e.g., hyperlinks, user links) between documents, or text spans in general, to enhance prediction performance for several NLP tasks. We think that a "network effect", similar to the one that took place in Information Retrieval (with the PageRank algorithm), could also positively impact NLP research. A few recent papers have already opened the way, for instance in attempting to exploit Twitter follower graph to improve sentiment classification [42].

Part of the challenge here will be to investigate how adequately and efficiently one can model these problems as instances of more general graph-based problems, such as node clustering/classification or link prediction discussed in the next sections. In a few cases, like text classification or sentiment analysis, graph modeling appears to be straightforward: nodes correspond to texts (and potentially users), and edges are given by relationships like hyperlinks, co-authorship, friendship, or thread membership. Unfortunately, modeling NLP problems as networks is not always that obvious. From the one hand, the right level of representation will probably vary depending on the task at hand: the nodes will be sentences, phrases, words, etc. From the other hand, the underlying graph will typically not be given a priori, which in turn raises the question of how we construct it. A preliminary discussion of the issue of optimal graph construction for semi-supervised learning in NLP is given in [21], [46]. We identify the issue of adaptive graph construction as an important scientific challenge for machine learning on graphs in general, and we will discuss it further in Section 3.3 .

As noted above, many NLP tasks have been recast as structured prediction problems, allowing to capture (some of the) output dependencies. How to best combine structured output and graph-based ML approaches is another challenge that we intend to address. We will initially investigate this question within a semi-supervised context, concentrating on graph regularization and graph propagation methods. Within such approaches, labels are typically binary or in a small finite set. Our objective is to explore how one propagates an exponential number of *structured labels* (like a sequence of tags or a dependency tree) through graphs. Recent attempts at blending structured output models with graph-based models are investigated in [43], [31]. Another related question that we will address in this context is how does one learn with *partial labels* (like partially specified tag sequence or tree) and use the graph structure to complete the output structure. This last question is very relevant to NLP problems where human annotations are costly; being able to learn from partial annotations could therefore allow for more targeted annotations and in turn reduced costs [33].

The NLP tasks we will mostly focus on are coreference resolution and entity linking, temporal structure prediction, and discourse parsing. These tasks will be envisioned in both document and cross-document settings, although we expect to exploit inter-document links either way. Choices for these particular tasks is guided by the fact that they are still open problems for the NLP community, they potentially have a high impact for industrial applications (like information retrieval, question answering, etc.), and we already have some expertise on these tasks in the team (see for instance [32], [28], [30]). As a midterm goal, we also plan to work on tasks more directly relating to micro-blogging, such as sentiment analysis and the automatic thread structuring of technical forums; the latter task is in fact an instance of rhetorical structure prediction [45].

We have already initiated some work on the coreference resolution with graph-based learning, by casting the problem as an instance of spectral clustering [30].

3.3. Adaptive Graph Construction

In most applications, edge weights are computed through a complex data modeling process and convey crucially important information for classifying nodes, making it possible to infer information related to each data sample even exploiting the graph topology solely. In fact, a widespread approach to several classification problems is to represent the data through an undirected weighted graph in which edge weights quantify the similarity between data points. This technique for coding input data has been applied to several domains, including classification of genomic data [40], face recognition [29], and text categorization [34].

In some cases, the full adjacency matrix is generated by employing suitable similarity functions chosen through a deep understanding of the problem structure. For example for the TF-IDF representation of documents, the affinity between pairs of samples is often estimated through the cosine measure or the χ^2 distance. After the generation of the full adjacency matrix, the second phase for obtaining the final graph consists in an edge sparsification/reweighting operation. Some of the edges of the clique obtained in the first step are pruned and the remaining ones can be reweighted to meet the specific requirements of the given classification problem. Constructing a graph with these methods obviously entails various kinds of loss of information. However, in problems like node classification, the use of graphs generated from several datasets can lead to an improvement in accuracy ([47], [22], [23]). Hence, the transformation of a dataset into a graph may, at least in some cases, partially remove various kinds of irregularities present in the original datasets, while keeping some of the most useful information for classifying the data samples. Moreover, it is often possible to accomplish classification tasks on the obtained graph using a running time remarkably lower than is needed by algorithms exploiting the initial datasets, and a suitable sparse graph representation can be seen as a compressed version of the original data. This holds even when input data are provided in a online/stream fashion, so that the resulting graph evolves over time.

In this project we will address the problem of adaptive graph construction towards several directions. The first one is about how to choose the best similarity measure given the objective learning task. This question is related to the question of metric and similarity learning ([24], [25]) which has not been considered in the context of graph-based learning. In the context of structured prediction, we will develop approaches where output structures are organized in graphs whose similarity is given by top-k outcomes of greedy algorithms.

A different way we envision adaptive graph construction is in the context of semi-supervised learning. Partial supervision can take various forms and an interesting and original setting is governed by two currently studied applications: detection of brain anomaly from connectome data and polls recommendation in marketing. Indeed, for these two applications, a partial knowledge of the information diffusion process can be observed while the network is unknown or only partially known. An objective is to construct (or complete) the network structure from some local diffusion information. The problem can be formalized as a graph construction problem from partially observed diffusion processes. It has been studied very recently in [36]. In our case, the originality comes either from the existence of different sources of observations or from the large impact of node contents in the network.

We will study how to combine graphs defined by networked data and graphs built from flat data to solve a given task. This is of major importance for information networks because, as said above, we will have to deal with multiple relations between entities (texts, spans of texts, ...) and also use textual data and vectorial data.

3.4. Prediction on Graphs and Scalability

As stated in the previous sections, graphs as complex objects provide a rich representation of data. Often enough the data is only partially available and the graph representation is very helpful in predicting the unobserved elements. We are interested in problems where the complete structure of the graph needs to be recovered and only a fraction of the links is observed. The link prediction problem falls into this category. We are also interested in the recommendation and link classification problems which can be seen as graphs

where the structure is complete but some labels on the links (weights or signs) are missing. Finally we are also interested in labeling the nodes of the graph, with class or cluster memberships or with a real value, provided that we have (some information about) the labels for some of the nodes.

The semi-supervised framework will be also considered. A midterm research plan is to study how graph regularization models help for structured prediction problems. This question will be studied in the context of NLP tasks, as noted in Section 3.2 , but we also plan to develop original machine learning algorithms that have a more general applicability. Inputs are networks whose nodes (texts) have to be labeled by structures. We assume that structures lie in some manifold and we want to study how labels can propagate in the network. One approach is to find a smooth labeling function corresponding to an harmonic function on both manifolds in input and output.

Scalability is one of the main issues in the design of new prediction algorithms working on networked data. It has gained more and more importance in recent years, because of the growing size of the most popular networked data that are now used by millions of people. In such contexts, learning algorithms whose computational complexity scales quadratically, or slower, in the number of considered data objects (usually nodes or edges, depending on the task) should be considered impractical.

These observations lead to the idea of using graph sparsification techniques in order to work on a part of the original network for getting results that can be easily extended and used for the whole original input. A sparsified version of the original graph can often be seen as a subset of the initial input, i.e. a suitably selected input subgraph which forms the training set (or, more in general, it is included in the training set). This holds even for the active setting. A simple example could be to find a spanning tree of the input graph, possibly using randomization techniques, with properties such that we are allowed to obtain interesting results for the initial graph dataset. We have started to explore this research direction for instance in [44].

At the level of mathematical foundations, the key issue to be addressed in the study of (large-scale) random networks also concerns the segmentation of network data into sets of independent and identically distributed observations. If we identify the data sample with the whole network, as it has been done in previous approaches [35], we typically end up with a set of observations (such as nodes or edges) which are highly interdependent and hence overly violate the classic i.i.d. assumption. In this case, the data scale can be so large and the range of correlations can be so wide, that the cost of taking into account the whole data and their dependencies is typically prohibitive. On the contrary, if we focus instead on a set of subgraphs independently drawn from a (virtually infinite) target network, we come up with a set of independent and identically distributed observations—namely the subgraphs themselves, where subgraph sampling is the underlying ergodic process [26]. Such an approach is one principled direction for giving novel statistical foundations to random network modeling. At the same time, because one shifts the focus from the whole network to a set of subgraphs, complexity issues can be restricted to the number of subgraphs and their size. The latter quantities can be controlled much more easily than the overall network size and dependence relationships, thus allowing to tackle scalability challenges through a radically redesigned approach.

Another way to tackle scalability problems is to exploit the inherent decentralized nature of very large graphs. Indeed, in many situations very large graphs are the abstract view of the digital activities of a very large set of users equipped with their own device. Nowadays, smartphones, tablets and even sensors have storage and computation power and gather a lot of data that serve to analytics, prediction, suggestion and personalized recommendation. Gathering all user data in large data centers is costly because it requires oversized infrastructures with huge energy consumption and large bandwidth networks. Even though cloud architectures can optimize such infrastructures, data concentration is also prone to security leaks, lost of privacy and data governance for end users. The alternative we have started to develop in Magnet is to devise decentralized, private and personalized machine learning algorithms so that they can be deployed in the personal devices. The key challenges are therefore to learn in a collaborative way in a network of learners and to preserve privacy and control on personal data.

3.5. Beyond Homophilic Relationships

In many cases, algorithms for solving node classification problems are driven by the following assumption: linked entities tend to be assigned to the same class. This assumption, in the context of social networks, is known as homophily ([27], [37]) and involves ties of every type, including friendship, work, marriage, age, gender, and so on. In social networks, homophily naturally implies that a set of individuals can be parted into subpopulations that are more cohesive. In fact, the presence of homogeneous groups sharing common interests is a key reason for affinity among interconnected individuals, which suggests that, in spite of its simplicity, this principle turns out to be very powerful for node classification problems in general networks.

Recently, however, researchers have started to consider networked data where connections may also carry a negative meaning. For instance, disapproval or distrust in social networks, negative endorsements on the Web. Although the introduction of signs on graph edges appears like a small change from standard weighted graphs, the resulting mathematical model, called signed graphs, has an unexpectedly rich additional complexity. For example, their spectral properties, which essentially all sophisticated node classification algorithms rely on, are different and less known than those of graphs. Signed graphs naturally lead to a specific inference problem that we have discussed in previous sections: link classification. This is the problem of predicting signs of links in a given graph. In online social networks, this may be viewed as a form of sentiment analysis, since we would like to semantically categorize the relationships between individuals.

Another way to go beyond homophily between entities will be studied using our recent model of hypergraphs with bipartite hyperedges [39]. A bipartite hyperedge connects two ends which are disjoint subsets of nodes. Bipartite hyperedges is a way to relate two collections of (possibly heterogeneous) entities represented by nodes. In the NLP setting, while hyperedges can be used to model bags of words, bipartite hyperedges are associated with relationships between bags of words. But each end of bipartite hyperedges is also a way to represent complex entities, gathering several attribute values (nodes) into hyperedges viewed as records. Our hypergraph notion naturally extends directed and undirected weighted graph. We have defined a spectral theory for this new class of hypergraphs and opened a way to smooth labeling on sets of nodes. The weighting scheme allows to weigh the participation of each node to the relationship modeled by bipartite hyperedges accordingly to an equilibrium condition. This condition provides a competition between nodes in hyperedges and allows interesting modeling properties that go beyond homophily and similarity over nodes (the theoretical analysis of our hypergraphs exhibits tight relationships with signed graphs). Following this competition idea, bipartite hyperedges are like matches between two teams and examples of applications are team creation. The basic tasks we are interested in are hyperedge classification, hyperedge prediction, node weight prediction. Finally, hypergraphs also represent a way to summarize or compress large graphs in which there exists highly connected couples of (large) subsets of nodes.

MOEX Project-Team

3. Research Program

3.1. Knowledge representation semantics

We work with semantically defined knowledge representation languages (like description logics, conceptual graphs and object-based languages). Their semantics is usually defined within model theory initially developed for logics.

We consider a language L as a set of syntactically defined expressions (often inductively defined by applying constructors over other expressions). A representation ($o \subseteq L$) is a set of such expressions. It may also be called an ontology. An interpretation function (I) is inductively defined over the structure of the language to a structure called the domain of interpretation (D). This expresses the construction of the "meaning" of an expression in function of its components. A formula is satisfied by an interpretation if it fulfills a condition (in general being interpreted over a particular subset of the domain). A model of a set of expressions is an interpretation satisfying all the expressions. A set of expressions is said consistent if it has at least one model, inconsistent otherwise. An expression (δ) is then a consequence of a set of expressions (o) if it is satisfied by all of their models (noted $o \models \delta$).

The languages dedicated to the semantic web (RDF and OWL) follow that approach. RDF is a knowledge representation language dedicated to the description of resources; OWL is designed for expressing ontologies: it describes concepts and relations that can be used within RDF.

A computer must determine if a particular expression (taken as a query, for instance) is the consequence of a set of axioms (a knowledge base). For that purpose, it uses programs, called provers, that can be based on the processing of a set of inference rules, on the construction of models or on procedural programming. These programs are able to deduce theorems (noted $o \vdash \delta$). They are said to be sound if they only find theorems which are indeed consequences and to be complete if they find all the consequences as theorems.

3.2. Data interlinking with link keys

Vast amounts of RDF data are made available on the web by various institutions providing overlapping information. To be fully exploited, different representations of the same object across various data sets, often using different ontologies, have to be identified. When different vocabularies are used for describing data, it is necessary to identify the concepts they define. This task is called ontology matching and its result is an alignment A, i.e. a set of correspondences $\langle e, r, e' \rangle$ relating entities e and e' of two different ontologies by a particular relation r (which may be equivalence, subsumption, disjointness, etc.) [4].

At the data level, data interlinking is the process of generating links identifying the same resource described in two data sets. Parallel to ontology matching, from two datasets (d and d') it generates a link set, L made of pairs of resource identifier.

We have introduced link keys [4], [1] which extend database keys in a way which is more adapted to RDF and deals with two data sets instead of a single relation. An example of a link key expression is:

$$\{\langle \mathsf{auteur}, \mathsf{creator} \rangle\}\{\langle \mathsf{titre}, \mathsf{title} \rangle\}\ \mathit{linkkey}\ \langle \mathsf{Livre}, \mathsf{Book} \rangle$$

stating that whenever an instance of the class Livre has the same values for the property auteur as an instance of class Book has for the property creator and they share at least one value for their property titre and title, then they denote the same entity. More precisely, a link key is a structure $\langle K^{eq}, K^{in}, C \rangle$ such that:

- K^{eq} and K^{in} are sets of pairs of property expressions;
- C is a pair of class expressions (or a correspondence).

Such a link key holds if and only if for any pair of resources belonging to the classes in correspondence such that the values of their property in K^{eq} are pairwise equal and the values of those in K^{in} pairwise intersect, the resources are the same. Link keys can then be used for finding equal individuals across two data sets and generating the corresponding owl:sameAs links. Link keys take into account the non functionality of RDF data and have to deal with non literal values. In particular, they may use arbitrary properties and class expressions. This renders their discovery and use difficult.

3.3. Experimental cultural knowledge evolution

Cultural evolution considers how culture spreads and evolves with human societies [21]. It applies an an idealised version of the theory of evolution to culture. In computer science, cultural evolution experiments are performed through multi-agent simulation: a society of agents adapts its culture through a precisely defined protocol [16]: agents perform repeatedly and randomly a specific task, called game, and their evolution is monitored. This aims at discovering experimentally the states that agents reach and the properties of these states.

Experimental cultural evolution has been successfully and convincingly applied to the evolution of natural languages [12], [23]. Agents play *language games* and adjust their vocabulary and grammar as soon as they are not able to communicate properly, i.e. they misuse a term or they do not behave in the expected way. It showed its capacity to model various such games in a systematic framework and to provide convincing explanations of linguistic phenomena. Such experiments have shown how agents can agree on a colour coding system or a grammatical case system.

Work has recently been developed for evolving alignments between ontologies. It can be used to repair alignments better than blind logical repair [19], to create alignments based on entity descriptions [13], to learn alignments from dialogues framed in interaction protocols [14], [18], or to correct alignments until no error remains [17][3] and to start with no alignment [2]. Each study provides new insights and opens perspectives.

We adapt this experimental strategy to knowledge representation [3]. Agents use their, shared or private, knowledge to play games and, in case of failure, they use adaptation operators to modify this knowledge. We monitor the evolution of agent knowledge with respect to their ability to perform the game (success rate) and with respect to the properties satisfied by the resulting knowledge itself. Such properties may, for instance, be:

- Agents converge to a common knowledge representation (a convergence property).
- Agents converge towards different but compatible (logically consistent) knowledge (a logical epistemic property), or towards closer knowledge (a metric epistemic property).
- That under the threat of a changing environment, agents that have operators that preserve diverse knowledge recover faster from the changes than those that have operators that converge towards a single representation (a differential property under environment change).

Our goal is to determine which operators are suitable for achieving desired properties in the context of a particular game.

ORPAILLEUR Project-Team

3. Research Program

3.1. Hybrid and Exploratory Knowledge Discovery

Keywords: knowledge discovery in databases, knowledge discovery in databases guided by domain knowledge, data mining, data exploration, formal concept analysis, classification, pattern mining, numerical methods in data mining.

Knowledge discovery in databases (KDD) aims at discovering intelligible and reusable patterns in possibly large databases. These patterns can then be interpreted as knowledge units to be reused in knowledge-based systems. From an operational point of view, the KDD process is based on three main steps: (i) selection and preparation of the data, (ii) data mining, (iii) interpretation of the discovered patterns. Moreover, the KDD process is iterative, interactive, and generally controlled by an expert of the data domain, called the analyst. The analyst selects and interprets a subset of the extracted units for obtaining knowledge units having a certain plausibility. In this view, KDD is an exploratory process similar to "exploratory data analysis".

The KDD process –as implemented in the Orpailleur team– is based on data mining methods which are either symbolic or numerical. Symbolic methods are based on pattern mining (e.g. mining frequent itemsets, association rules, sequences...), Formal Concept Analysis (FCA) and extensions such as Pattern Structures and Relational Concept Analysis (RCA), and redescription mining. Numerical methods are based on Random Forests, Support Vector Machines (SVM), Neural Networks, and probabilistic approaches such as second-order Hidden Markov Models (HMM). Moreover, for being able to deal with complex data, numerical data mining methods can be associated with symbolic methods, for improving applicability and efficiency of knowledge discovery. This is particularly true in classification, where supervised and unsupervised approaches may be combined with benefits.

A main operation in the research work of Orpailleur is "classification", which is a polymorphic process involved in modeling, mining, representing, and reasoning tasks. In this way, domain knowledge, when available, can improve and guide the KDD process, materializing the idea of *Knowledge Discovery guided by Domain Knowledge* or KDDK. In KDDK, domain knowledge plays a role at each step of KDD: the discovered patterns can be interpreted as knowledge units and reused for problem-solving activities in knowledge systems, implementing the exploratory process "mining, interpreting, modeling, representing, and reasoning". Then knowledge discovery can be considered as a key task in knowledge engineering (KE), having an impact in various semantic activities, e.g. information retrieval, recommendation, and ontology engineering. In addition, if knowledge discovery can feed knowledge-based systems, in turn, domain knowledge can be used to support the knowledge discovery process.

Finally, life sciences, i.e. agronomy, biology, chemistry, and medicine, are application domains where the Orpailleur team has a very rich experience. The team intends to keep and to extend this experience, paying also more attention to the impact of knowledge discovery in the real world. This should lead to the design of green (sustainable), explainable, and fair data mining systems.

3.2. Text Mining

Keywords: text mining, knowledge discovery from texts, text classification, annotation, ontology engineering from texts.

The objective of a text mining process is to extract useful knowledge units from large collections of texts [71]. The text mining process shows specific characteristics due to the fact that texts are complex objects written in natural language. The information in a text is expressed in an informal way, following linguistic rules, making text mining a difficult task. A text mining process has to take into account –as much as possible– paraphrases, ambiguities, specialized vocabulary and terminology. This is why the preparation of texts for text mining is usually dependent on linguistic resources and methods.

From a knowledge discovery perspective, text mining aims at extracting "interesting units" (nouns and relations) from texts with the help of domain knowledge encoded within a knowledge base. The process is roughly similar for text annotation. Text mining is especially useful in the context of semantic web for ontology engineering. In the Orpailleur team, we work on the mining of real-world texts in application domains such as biology and medicine, using numerical and symbolic data mining methods. Accordingly, the text mining process may be involved in a loop used to enrich and to extend linguistic resources. In turn, linguistic and ontological resources can be exploited to guide a "knowledge-based text mining process".

3.3. Knowledge Systems and Web of Data

Keywords: knowledge engineering, web of data, semantic web, ontology, description logics, classification-based reasoning, case-based reasoning, information retrieval, recommendation.

The web of data constitutes a good platform for experimenting ideas on knowledge engineering (KE) and knowledge discovery. A software agent may be able to read, understand, and manipulate information on the web, if and only if the knowledge necessary for achieving those tasks is available. This is why domain knowledge and ontologies are of main importance. OWL ("Web Ontology Language" https://www.w3.org/OWL/) is based on description logics (DLs [72]) and is the representation language commonly used for designing ontologies. In OWL, knowledge units are represented by classes having properties and instances. Concepts are organized within a partially ordered set based on a subsumption relation, and the inference services are based on subsumption and classification.

Actually, there are many interconnections between concept lattices in FCA and ontologies, e.g. the partial order underlying an ontology can be supported by a concept lattice. Moreover, a pair of implications within a concept lattice can provide a possible materialization of a concept definition in an ontology. In this way, we study how the web of data, considered as a set of knowledge sources, e.g. DBpedia, Wikipedia, Yago, Freebase, can be mined for guiding the design of a knowledge base, and further, how knowledge discovery techniques can be applied for allowing a better usage of the web of data, e.g. Linked Open Data (LOD) classification and completion.

Then, a part of the research work in Knowledge Engineering is oriented towards knowledge discovery in the web of data, as, with the increased interest in machine processable data, more and more data is now published in RDF (Resource Description Framework) format. Particularly, we are interested in the completeness of the data and their potential to provide concept definitions in terms of necessary and sufficient conditions. We have proposed algorithms based on FCA and Redescription Mining which allow data exploration as well as the discovery of definition (bidirectional implication rules).

PETRUS Project-Team

3. Research Program

3.1. Research Program

To tackle the challenge introduced above, we identify three main lines of research:

- (Axis 1) Personal cloud server architectures. Based on the intuition that user control, security and privacy are key properties in the definition of trusted personal cloud solutions, our objective is to propose new architectures (encompassing both software and hardware aspects) for secure personal cloud data management and formally prove important bricks of the architecture. We also focus in this axis on administration models and their enforcement in relation to the architecture of the system, so that the exclusive control of a non expert individual can be ensured.
- (Axis 2) Global query evaluation. The goal of this line of research is to provide capabilities for crossing data belonging to multiple individuals (e.g., performing statistical queries over personal data, computing queries on social graphs or organizing participatory data collection) in a fully decentralized setting while providing strong and personalized privacy guarantees. This means proposing new secure distributed database indexing models and query processing strategies. In addition, we concentrate on locally ensuring to each participant the good behaviour of the processing, such that no collective results can be produced if privacy conditions are not respected by other participants.
- (Axis 3) Economic, legal and societal issues. This research axis is more transverse and entails multidisciplinary research, addressing the links between economic, legal, societal and technological aspects. We will follow here a multi-disciplinary approach based on a 3-step methodology: i) identifying important common issues related to privacy and to the exploitation of personal data; ii) characterizing their dimensions in all relevant disciplines and jointly study their entanglement; iii) validating the proposed analysis, models and trade-offs thanks to in vivo experiments.

These contributions will also rely on tools (algorithms, protocols, proofs, etc.) from other communities, namely security (cryptography, secure multiparty computations, formal methods, differential privacy, etc.) and distributed systems (distributed hash tables, gossip protocols, etc.). Beyond the research actions, we structure our software activity around a single common platform (rather than isolated demonstrators), integrating our main research contributions, called PlugDB. This platform is the cornerstone to help validating our research results through accurate performance measurements on a real platform, a common practice in the DB community, and target the best conferences. It is also a strong vector to federate the team, simplify the bootstrapping of new PhD or master students, conduct multi-disciplinary research and open the way to industrial collaborations and technological transfers.

TYREX Project-Team

3. Research Program

3.1. Foundations for Data Manipulation Analysis: Logics and Type Systems

We develop methods for the static analysis of queries based on logical decision procedures. Static analysis can be used to optimize runtime performance by compile-time automated modification of the code. For example, queries can be substituted by more efficient — yet equivalent — variants. The query containment problem has been a central point of research for major query languages due to its vital role in query optimization. Query containment is defined as determining if the result of one query is included in the result of another one for any dataset. We explore techniques for deciding query containment for expressive languages for querying richly structured data such as knowledge graphs. One major scientific difficulty here consists in dealing with problems close to the frontier of decidability, and therefore in finding useful trade-offs between programming expressivity, complexity, succinctness, algorithmic techniques and effective implementations. We also investigate type systems and type-checking methods for the analysis of the manipulations of structured data.

3.2. Algebraic Foundations for Query Optimization and Code Synthesis

We consider intermediate languages based on algebraic foundations for the representation, characterization, transformations and compilation of queries. We investigate extensions of the relational algebra for optimizing expressive queries, and in particular recursive queries. We explore monads and in particular monad comprehensions and monoid calculus for the generation of efficient and scalable code on big data frameworks. When transforming and optimizing algebraic terms, we rely on cost-based searches of equivalent terms. We thus develop cost models whose purpose is to estimate the time, space and network costs of query evaluation. One difficulty is to estimate these costs in architectures where data and computations are distributed, and where the modeling of data transfers is essential.

VALDA Project-Team

3. Research Program

3.1. Scientific Foundations

We now detail some of the scientific foundations of our research on complex data management. This is the occasion to review connections between data management, especially on complex data as is the focus of Valda, with related research areas.

3.1.1. Complexity & Logic

Data management has been connected to logic since the advent of the relational model as main representation system for real-world data, and of first-order logic as the logical core of database querying languages [37]. Since these early developments, logic has also been successfully used to capture a large variety of query modes, such as data aggregation [63], recursive queries (Datalog), or querying of XML databases [46]. Logical formalisms facilitate reasoning about the expressiveness of a query language or about its complexity.

The main problem of interest in data management is that of query evaluation, i.e., computing the results of a query over a database. The complexity of this problem has far-reaching consequences. For example, it is because first-order logic is in the AC_0 complexity class that evaluation of SQL queries can be parallelized efficiently. It is usual [74] in data management to distinguish *data complexity*, where the query is considered to be fixed, from *combined complexity*, where both the query and the data are considered to be part of the input. Thus, though conjunctive queries, corresponding to a simple SELECT-FROM-WHERE fragment of SQL, have PTIME data complexity, they are NP-hard in combined complexity. Making this distinction is important, because data is often far larger (up to the order of terabytes) than queries (rarely more than a few hundred bytes). Beyond simple query evaluation, a central question in data management remains that of complexity; tools from algorithm analysis, and complexity theory can be used to pinpoint the tractability frontier of data management tasks.

3.1.2. Automata Theory

Automata theory and formal languages arise as important components of the study of many data management tasks: in temporal databases [36], queries, expressed in temporal logics, can often by compiled to automata; in graph databases [42], queries are naturally given as automata; typical query and schema languages for XML databases such as XPath and XML Schema can be compiled to tree automata [67], or for more complex languages to data tree automata[4]. Another reason of the importance of automata theory, and tree automata in particular, comes from Courcelle's results [50] that show that very expressive queries (from the language of monadic second-order language) can be evaluated as tree automata over *tree decompositions* of the original databases, yielding linear-time algorithms (in data complexity) for a wide variety of applications.

3.1.3. Verification

Complex data management also has connections to verification and static analysis. Besides query evaluation, a central problem in data management is that of deciding whether two queries are *equivalent* [37]. This is critical for query optimization, in order to determine if the rewriting of a query, maybe cheaper to evaluate, will return the same result as the original query. Equivalence can easily be seen to be an instance of the problem of (non-)satisfiability: $q \equiv q'$ if and only if $(q \wedge \neg q') \vee (\neg q \wedge q')$ is not satisfiable. In other words, some aspects of query optimization are static analysis issues. Verification is also a critical part of any database application where it is important to ensure that some property will never (or always) arise [48].

3.1.4. Workflows

The orchestration of distributed activities (under the responsibility of a conductor) and their choreography (when they are fully autonomous) are complex issues that are essential for a wide range of data management applications including notably, e-commerce systems, business processes, health-care and scientific workflows. The difficulty is to guarantee consistency or more generally, quality of service, and to statically verify critical properties of the system. Different approaches to workflow specifications exist: automata-based, logic-based, or predicate-based control of function calls [34].

3.1.5. Probability & Provenance

To deal with the uncertainty attached to data, proper models need to be used (such as attaching *provenance* information to data items and viewing the whole database as being *probabilistic*) and practical methods and systems need to be developed to both reliably estimate the uncertainty in data items and properly manage provenance and uncertainty information throughout a long, complex system.

The simplest model of data uncertainty is the NULLs of SQL databases, also called Codd tables [37]. This representation system is too basic for any complex task, and has the major inconvenient of not being closed under even simple queries or updates. A solution to this has been proposed in the form of *conditional tables* [61] where every tuple is annotated with a Boolean formula over independent Boolean random events. This model has been recognized as foundational and extended in two different directions: to more expressive models of *provenance* than what Boolean functions capture, through a semiring formalism [57], and to a probabilistic formalism by assigning independent probabilities to the Boolean events [58]. These two extensions form the basis of modern provenance and probability management, subsuming in a large way previous works [49], [43]. Research in the past ten years has focused on a better understanding of the tractability of query answering with provenance and probabilistic annotations, in a variety of specializations of this framework [72] [62], [40].

3.1.6. Machine Learning

Statistical machine learning, and its applications to data mining and data analytics, is a major foundation of data management research. A large variety of research areas in complex data management, such as wrapper induction [68], crowdsourcing [41], focused crawling [56], or automatic database tuning [44] critically rely on machine learning techniques, such as classification [60], probabilistic models [55], or reinforcement learning [73].

Machine learning is also a rich source of complex data management problems: thus, the probabilities produced by a conditional random field [65] system result in probabilistic annotations that need to be properly modeled, stored, and queried.

Finally, complex data management also brings new twists to some classical machine learning problems. Consider for instance the area of *active learning* [70], a subfield of machine learning concerned with how to optimally use a (costly) oracle, in an interactive manner, to label training data that will be used to build a learning model, e.g., a classifier. In most of the active learning literature, the cost model is very basic (uniform or fixed-value costs), though some works [69] consider more realistic costs. Also, oracles are usually assumed to be perfect with only a few exceptions [53]. These assumptions usually break when applied to complex data management problems on real-world data, such as crowdsourcing.

3.2. Research Directions

At the beginning of the Valda team, the project was to focus on the following directions:

- foundational aspects of data management, in particular related to query enumeration and reasoning on data, especially regarding security issues;
- implementation of provenance and uncertainty management, real-world applications, other aspects of uncertainty and incompleteness, in particular dynamic;
- development of personal information management systems, integration of machine learning techniques.

We believe the first two directions have been followed in a satisfactory manner. The focus on personal information management has not been kept for various organizational reasons, however, but the third axis of the project is reoriented to more general aspects of Web data management.

New permanent arrivals in the group since its creation have impacted its research directions in the following manner:

- Camille BOURGAUX and Michaël THOMAZO are both specialists of knowledge representation and formal aspects of knowledge bases, which is an expertise that did not exist in the group. They are also both interested in, and have started working on aspects related to connecting their research with database theory, and investigating aspects of uncertainty and incompleteness in their research. This will lead to more work on knowledge representation and symbolic AI aspects, while keeping the focus of Valda on foundations of data management and uncertainty.
- Olivier CAPPÉ is a specialist in statistics and machine learning, in particular multi-armed bandits and reinforcement learning. He is also interested in applications of these learning techniques to data management problems. His arrival in the group therefore complements the expertise of other researchers, and will lead to more work on machine learning issues.
- Leonid LIBKIN is a specialist of database theory, of incomplete data management, and has a line of current research on graph data management. His profile fits very well with the original orientation of the Valda project.

We intend to keep producing leading research on the foundations of data management. Generally speaking, the goal is to investigate the borders of feasibility of various tasks. For instance, what are the assumptions on data that allow for computable problems? When is it not possible at all? When can we hope for efficient query answering, when is it hopeless? This is a problem of theoretical nature which is necessary for understanding the limit of the methods and driving research towards the scenarios where positive results may be obtainable. Only when we have understood the limitation of different methods and have many examples where this is possible, we can hope to design a solid foundation that allowing for a good trade-off between what can be done (needs from the users) and what can be achieved (limitation from the system).

Similarly, we will continue our work, both foundational and practical, on various aspects of provenance and uncertainty management. One overall long-term goal is to reach a full understanding of the interactions between query evaluation or other broader data management tasks and uncertain and annotated data models. We would in particular want to go towards a full classification of tractable (typically polynomial-time) and intractable (typically NP-hard for decision problems, or #P-hard for probability evaluation) tasks, extending and connecting the query-based dichotomy [51] on probabilistic query evaluation with the instance-based one of [39], [40]. Another long-term goal is to consider more dynamic scenarios than what has been considered so far in the uncertain data management literature: when following a workflow, or when interacting with intensional data sources, how to properly represent and update uncertainty annotations that are associated with data. This is critical for many complex data management scenarios where one has to maintain a probabilistic current knowledge of the world, while obtaining new knowledge by posing queries and accessing data sources. Such intensional tasks requires minimizing jointly data uncertainty and cost to data access.

As application area, in addition to the historical focus on personal information management which is now less stressed, we target Web data (Web pages, the semantic Web, social networks, the deep Web, crowdsourcing platforms, etc.).

We aim at keeping a delicate balance between theoretical, foundational research, and systems research, including development and implementation. This is a difficult balance to find, especially since most Valda researchers have a tendency to favor theoretical work, but we believe it is also one of the strengths of the team.

WIMMICS Project-Team

3. Research Program

3.1. Users Modeling and Designing Interaction on the Web

Wimmics focuses on interactions of ordinary users with ontology-based knowledge systems, with a preference for semantic Web formalisms and Web 2.0 applications. We specialize interaction design and evaluation methods to Web application tasks such as searching, browsing, contributing or protecting data. The team is especially interested in using semantics in assisting the interactions. We propose knowledge graph representations and algorithms to support interaction adaptation, for instance for context-awareness or intelligent interactions with machine. We propose and evaluate Web-based visualization techniques for linked data, querying, reasoning, explaining and justifying. Wimmics also integrates natural language processing approaches to support natural language based interactions. We rely on cognitive studies to build models of the system, the user and the interactions between users through the system, in order to support and improve these interactions. We extend the user modeling technique known as *Personas* where user models are represented as specific, individual humans. *Personas* are derived from significant behavior patterns (i.e., sets of behavioral variables) elicited from interviews with and observations of users (and sometimes customers) of the future product. Our user models specialize *Personas* approaches to include aspects appropriate to Web applications. Wimmics also extends user models to capture very different aspects (e.g. emotional states).

3.2. Communities and Social Interactions Analysis

The domain of social network analysis is a whole research domain in itself and Wimmics targets what can be done with typed graphs, knowledge representations and social models. We also focus on the specificity of social Web and semantic Web applications and in bridging and combining the different social Web data structures and semantic Web formalisms. Beyond the individual user models, we rely on social studies to build models of the communities, their vocabularies, activities and protocols in order to identify where and when formal semantics is useful. We propose models of collectives of users and of their collaborative functioning extending the collaboration personas and methods to assess the quality of coordination interactions and the quality of coordination artifacts. We extend and compare community detection algorithms to identify and label communities of interest with the topics they share. We propose mixed representations containing social semantic representations (e.g. folksonomies) and formal semantic representations (e.g. ontologies) and propose operations that allow us to couple them and exchange knowledge between them. Moving to social interaction we develop models and algorithms to mine and integrate different yet linked aspects of social media contributions (opinions, arguments and emotions) relying in particular on natural language processing and argumentation theory. To complement the study of communities we rely on multi-agent systems to simulate and study social behaviors. Finally we also rely on Web 2.0 principles to provide and evaluate social Web applications.

3.3. Vocabularies, Semantic Web and Linked Data Based Knowledge Representation and Artificial Intelligence Formalisms on the Web

For all the models we identified in the previous sections, we rely on and evaluate knowledge representation methodologies and theories, in particular ontology-based modeling. We also propose models and formalisms to capture and merge representations of different levels of semantics (e.g. formal ontologies and social folksonomies). The important point is to allow us to capture those structures precisely and flexibly and yet create as many links as possible between these different objects. We propose vocabularies and semantic Web formalizations for all the aspects that we model and we consider and study extensions of these formalisms when needed. The results have all in common to pursue the representation and publication of our models

as linked data. We also contribute to the transformation and linking of existing resources (informal models, databases, texts, etc.) to be published on the Semantic Web and as Linked Data. Examples of aspects we formalize include: user profiles, social relations, linguistic knowledge, business processes, derivation rules, temporal descriptions, explanations, presentation conditions, access rights, uncertainty, emotional states, licenses, learning resources, etc. At a more conceptual level we also work on modeling the Web architecture with philosophical tools so as to give a realistic account of identity and reference and to better understand the whole context of our research and its conceptual cornerstones.

3.4. Artificial Intelligence Processing: Learning, Analyzing and Reasoning on Heterogeneous Semantic Graphs

One of the characteristics of Wimmics is to rely on graph formalisms unified in an abstract graph model and operators unified in an abstract graph machine to formalize and process semantic Web data, Web resources, services metadata and social Web data. In particular Corese, the core software of Wimmics, maintains and implements that abstraction. We propose algorithms to process the mixed representations of the previous section. In particular we are interested in allowing cross-enrichment between them and in exploiting the life cycle and specificity of each one to foster the life-cycles of the others. Our results all have in common to pursue analyzing and reasoning on heterogeneous semantic graphs issued from social and semantic Web applications. Many approaches emphasize the logical aspect of the problem especially because logics are close to computer languages. We defend that the graph nature of Linked Data on the Web and the large variety of types of links that compose them call for typed graphs models. We believe the relational dimension is of paramount importance in these representations and we propose to consider all these representations as fragments of a typed graph formalism directly built above the Semantic Web formalisms. Our choice of a graph based programming approach for the semantic and social Web and of a focus on one graph based formalism is also an efficient way to support interoperability, genericity, uniformity and reuse.

ZENITH Project-Team

3. Research Program

3.1. Distributed Data Management

Data management is concerned with the storage, organization, retrieval and manipulation of data of all kinds, from small and simple to very large and complex. It has become a major domain of computer science, with a large international research community and a strong industry. Continuous technology transfer from research to industry has led to the development of powerful DBMS, now at the heart of any information system, and of advanced data management capabilities in many kinds of software products (search engines, application servers, document systems, etc.).

To deal with the massive scale of scientific data, we exploit large-scale distributed systems, with the objective of making distribution transparent to the users and applications. Thus, we capitalize on the principles of large-scale distributed systems such as clusters, peer-to-peer (P2P) and cloud.

Data management in distributed systems has been traditionally achieved by distributed database systems which enable users to transparently access and update several databases in a network using a high-level query language (e.g. SQL). Transparency is achieved through a global schema which hides the local databases' heterogeneity. In its simplest form, a distributed database system supports a global schema and implements distributed database techniques (query processing, transaction management, consistency management, etc.). This approach has proved to be effective for applications that can benefit from centralized control and full-fledge database capabilities, e.g. information systems. However, it cannot scale up to more than tens of databases.

Parallel database systems extend the distributed database approach to improve performance (transaction throughput or query response time) by exploiting database partitioning using a multiprocessor or cluster system. Although data integration systems and parallel database systems can scale up to hundreds of data sources or database partitions, they still rely on a centralized global schema and strong assumptions about the network.

In contrast, peer-to-peer (P2P) systems adopt a completely decentralized approach to data sharing. By distributing data storage and processing across autonomous peers in the network, they can scale without the need for powerful servers. P2P systems typically have millions of users sharing petabytes of data over the Internet. Although very useful, these systems are quite simple (e.g. file sharing), support limited functions (e.g. keyword search) and use simple techniques (e.g. resource location by flooding) which have performance problems. A P2P solution is well-suited to support the collaborative nature of scientific applications as it provides scalability, dynamicity, autonomy and decentralized control. Peers can be the participants or organizations involved in collaboration and may share data and applications while keeping full control over their (local) data sources. But for very-large scale scientific data analysis, we believe cloud computing (see next section), is the right approach as it can provide virtually infinite computing, storage and networking resources. However, current cloud architectures are proprietary, ad-hoc, and may deprive users of the control of their own data. Thus, we postulate that a hybrid P2P/cloud architecture is more appropriate for scientific data management, by combining the best of both approaches. In particular, it will enable the clean integration of the users' own computational resources with different clouds.

3.2. Big Data

Big data (like its relative, data science) has become a buzz word, with different meanings depending on your perspective, e.g. 100 terabytes is big for a transaction processing system, but small for a web search engine. It is also a moving target, as shown by two landmarks in DBMS products: the Teradata database machine in the 1980's and the Oracle Exadata database machine in 2010.

Although big data has been around for a long time, it is now more important than ever. We can see overwhelming amounts of data generated by all kinds of devices, networks and programs, e.g. sensors, mobile devices, connected objects (IoT), social networks, computer simulations, satellites, radiotelescopes, etc. Storage capacity has doubled every 3 years since 1980 with prices steadily going down (e.g. 1 Gigabyte of Hard Disk Drive for: 1M$ in 1982, 1K$ in 1995, 0.02$ in 2015), making it affordable to keep more data around. And massive data can produce high-value information and knowledge, which is critical for data analysis, decision support, forecasting, business intelligence, research, (data-intensive) science, etc.

The problem of big data has three main dimensions, quoted as the three big V's:

- Volume: refers to massive amounts of data, making it hard to store, manage, and analyze (big analytics);
- Velocity: refers to continuous data streams being produced, making it hard to perform online processing and analysis;
- Variety: refers to different data formats, different semantics, uncertain data, multiscale data, etc., making it hard to integrate and analyze.

There are also other V's such as: validity (is the data correct and accurate?); veracity (are the results meaningful?); volatility (how long do you need to store this data?).

Many different big data management solutions have been designed, primarily for the cloud, as cloud and big data are synergistic. They typically trade consistency for scalability, simplicity and flexibility, hence the new term Data-Intensive Scalable Computing (DISC). Examples of DISC systems include data processing frameworks (e.g. Hadoop MapReduce, Apache Spark, Pregel), file systems (e.g. Google GFS, HDFS), NoSQL systems (Google BigTable, Hbase, MongoDB), NewSQL systems (Google F1, CockroachDB, LeanXcale). In Zenith, we exploit or extend DISC technologies to fit our needs for scientific workflow management and scalable data analysis.

3.3. Data Integration

Scientists can rely on web tools to quickly share their data and/or knowledge. Therefore, when performing a given study, a scientist would typically need to access and integrate data from many data sources (including public databases). Data integration can be either physical or logical. In the former, the source data are integrated and materialized in a data warehouse. In logical integration, the integrated data are not materialized, but accessed indirectly through a global (or mediated) schema using a data integration system. These two approaches have different trade-offs, e.g. efficient analytics but only on historical data for data warehousing versus real-time access to data sources for data integration systems (e.g. web price comparators).

In both cases, to understand a data source content, metadata (data that describe the data) is crucial. Metadata can be initially provided by the data publisher to describe the data structure (e.g. schema), data semantics based on ontologies (that provide a formal representation of the domain knowledge) and other useful information about data provenance (publisher, tools, methods, etc.). Scientific metadata is very heterogeneous, in particular because of the autonomy of the underlying data sources, which leads to a large variety of models and formats. Thus, it is necessary to identify semantic correspondences between the metadata of the related data sources. This requires the matching of the heterogeneous metadata, by discovering semantic correspondences between ontologies, and the annotation of data sources using ontologies. In Zenith, we rely on semantic web techniques (e.g. RDF and SparkQL) to perform these tasks and deal with high numbers of data sources.

Scientific workflow management systems (SWfMS) are also useful for data integration. They allow scientists to describe and execute complex scientific activities, by automating data derivation processes, and supporting various functions such as provenance management, queries, reuse, etc. Some workflow activities may access or produce huge amounts of distributed data. This requires using distributed and parallel execution environments. However, existing workflow management systems have limited support for data parallelism. In Zenith, we use an algebraic approach to describe data-intensive workflows and exploit parallelism.

3.4. Data Analytics

Data analytics refers to a set of techniques to draw conclusions through data examination. It involves data mining, statistics, and data management, and is applied to categorical and continuous data. In the Zenith team, we are interested in both of these data types. Categorical data designates a set of data that can be described as "check boxes". It can be names, products, items, towns, etc. A common illustration is the market basket data, where each item bought by a client is recorded and the set of items is the basket. The typical data mining problems with this kind of data are:

- **Frequent itemsets and association rules**. In this case, the data is usually a table with a high number of rows and the data mining algorithm extracts correlations between column values. A typical example of frequent itemset from a sensor network in a smart building would say that "in 20% rooms, the door is closed, the room is empty, and lights are on."
- **Frequent sequential pattern extraction.** This problem is very similar to frequent itemset discovery but considering the order between. In the smart building example, a frequent sequence could say that "in 40% of rooms, lights are on at time i, the room is empty at time $i + j$ and the door is closed at time $i + j + k$".
- **Clustering**. The goal of clustering is to group together similar data while ensuring that dissimilar data will not be in the same cluster. In our example of smart buildings, we could find clusters of rooms, where offices will be in one category and copy machine rooms in another because of their differences (hours of people presence, number of times lights are turned on/off, etc.).

Continuous data are numeric records that can have an infinite number of values between any two values. A temperature value or a timestamp are examples of such data. They are involved in a widely used type of data known as time series: a series of values, ordered by time, and giving a measure, e.g. coming from a sensor. There is a large number of problems that can apply to this kind of data, including:

- **Indexing and retrieval.** The goal, here, is usually to find, given a query q and a time series dataset D, the records of D that are most similar to q. This may involve any transformation of D by means of an index or an alternative representation for faster execution.
- **Pattern and outlier detection.** The discovery of recurrent patterns or atypical sub-windows in a time series has applications in finance, industrial manufacture or seismology, to name a few. It calls for techniques that avoid pairwise comparisons of all the sub-windows, which would lead to prohibitive response times.
- **Clustering**. The goal is the same as categorical data clustering: group similar time series and separate dissimilar ones.

One main problem in data analytics is to deal with data streams. Existing methods have been designed for very large data sets where complex algorithms from artificial intelligence were not efficient because of data size. However, we now must deal with data streams, sequences of data events arriving at high rate, where traditional data analytics techniques cannot complete in real-time, given the infinite data size. In order to extract knowledge from data streams, the data mining community has investigated approximation methods that could yield good result quality.

3.5. High dimensional data processing and search

High dimensionality is inherent in applications involving images, audio and text as well as in many scientific applications involving raster data or high-throughput data. Because of the *dimensionality curse*, technologies for processing and analyzing such data cannot rely on traditional relational DBMS or data mining methods. It rather requires to employ machine learning methods such as dimensionality reduction, representation learning or random projection. The activity of Zenith in this domain focuses on methods that permit data processing and search at scale, in particular in the presence of strong uncertainty and/or ambiguity. Actually, while small datasets are often characterized by a careful collection process, massive amounts of data often come with outliers and spurrious items, because it appears impossible to guarantee faultless collection at massive

bandwidth. Another source of noise is often the sensor itself, that may be of low quality but of high sampling rate, or even the actual content, e.g. in cultural heritage applications when historical content appears seriously damaged by time. To attack these difficult problems, we focus on the following research topics:

- **Uncertainty estimation**. Items in massive datasets may either be uncertain, e.g. for automatically annotated data as in image analysis, or be more or less severely corrupted by noise, e.g. in noisy audio recordings or in the presence of faulty sensors. In both cases, the concept of *uncertainty* is central for the end-user to exploit the content. In this context, we use probability theory to quantify uncertainty and propose machine learning algorithms that may operate robustly, or at least assess the quality of their output. This vast topic of research is guided by large-scale applications (both data search and data denoising), and our research is oriented towards computationally effective methods.
- **Deep neural networks**. A major breakthrough in machine learning performance has been the advent of deep neural networks, which are characterized by high numbers (millions) of parameters and scalable learning procedures. We are striving towards original architectures and methods that are theoretically grounded and offer state-of-the-art performance for data search and processing. The specific challenges we investigate are: very high dimensionality for static data and very long-term dependency for temporal data, both in the case of possibly strong uncertainty or ambiguity (e.g. hundreds of thousands of classes).
- **Community service**. Research in machine learning is guided by applications. In Zenith, two main communities are targeted: botany, and digital humanities. In both cases, our observation is that significant breakthroughs may be achieved by connecting these communities to machine learning researchers. This may be achieved through wording application-specific problems in classical machine learning parlance. Thus, the team is actively involved in the organization of international evaluation campaigns that allow machine learning researchers to propose new methods while solving important application problems.

ALICE Team

3. Research Program

3.1. Point clouds

Currently, transforming the raw point cloud into a triangular mesh is a long pipeline involving disparate geometry processing algorithms:

- *Point pre-processing:* colorization, filtering to remove unwanted background, first noise reduction along acquisition viewpoint;
- *Registration:* cloud-to-cloud alignment, filtering of remaining noise, registration refinement;
- *Mesh generation:* triangular mesh from the complete point cloud, re-meshing, smoothing.

The output of this pipeline is a locally structured model which is used in downstream mesh analysis methods such as feature extraction, segmentation in meaningful parts or building CAD models.

It is well known that point cloud data contains measurement errors due to factors related to the external environment and to the measurement system itself [37], [32], [20]. These errors propagate through all processing steps: pre-processing, registration and mesh generation. Even worse, the heterogeneous nature of different processing steps makes it extremely difficult to know *how* these errors propagate through the pipeline. To give an example, for cloud-to-cloud alignment it is necessary to estimate normals. However, the normals are forgotten in the point cloud produced by the registration stage. Later on, when triangulating the cloud, the normals are re-estimated on the modified data, thus introducing uncontrollable errors.

We plan to develop new reconstruction, meshing and re-meshing algorithms, with a specific focus on the accuracy and resistance to all defects present in the input raw data. We think that pervasive treatment of uncertainty is the missing ingredient to achieve this goal. We plan to rethink the pipeline with the position uncertainty maintained during the whole process. Input points can be considered either as error ellipsoids [41] or as probability measures [27]. In a nutshell, our idea is to start by computing an error ellipsoid [43], [29] for each point of the raw data, and then to cumulate the errors (approximations) committed at each step of the processing pipeline while building the mesh. In this way, the final users will be able to take the uncertainty knowledge into account and rely on this confidence measure for further analysis and simulations. Quantifying uncertainty for reconstruction algorithms, and propagating them from input data to high-level geometry processing algorithms has never been considered before, possibly due to the very different methodologies of the approaches involved. At the very beginning we will re-implement the entire pipeline, and then attack the weak links through all three reconstruction stages.

3.2. Parameterizations

One of the favorite tools we use in our team are parameterizations. They provide a very powerful way to reveal structures on objects. The most omnipresent application of parameterizations is texture mapping: texture maps provide a way to represent in 2D (on the map) information related to a surface. Once the surface is equipped with a map, we can do much more than a mere coloring of the surface: we can approximate geodesics, edit the mesh directly in 2D or transfer information from one mesh to another.

Parameterizations constitute a family of methods that involve optimizing an objective function, subject to a set of constraints (equality, inequality, being integer, etc.). Computing the exact solution to such problems is beyond any hope, therefore approximations are the only resort. This raises a number of problems, such as the minimization of highly nonlinear functions and the definition of direction fields topology, without forgetting the robustness of the software that puts all this into practice.

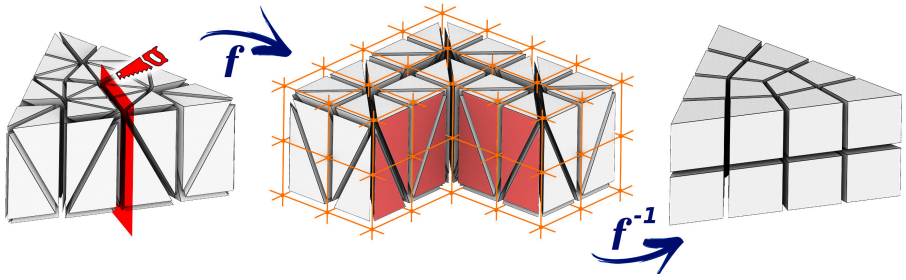


Figure 2. Hex-remeshing via global parameterization. ***Left:*** *Input tetrahedral mesh. To allow for a singular edge in the center, the mesh is cut open along the red plane.* ***Middle:*** *Mesh in parametric space.* ***Right:*** *Output mesh defined by parameterization.*

We are particularly interested in a specific instance of parameterization: hexahedral meshing. The idea [4] is to build a transformation f from the domain to a parametric space, where the deformed domain can be meshed by a regular grid. The inverse transformation f^{-1} applied to this grid produces the hexahedral mesh of the domain, aligned with the boundary of the object. The strength of this approach is that the transformation may admit some discontinuities. Let us show an example: we start from a tetrahedral mesh (Figure 2 , left) and we want deform it in a way that its boundary is aligned with the integer grid. To allow for a singular edge in the output (the valency 3 edge, Figure 2 , right), the input mesh is cut open along the highlighted faces and the central edge is mapped onto an integer grid line (Figure 2 , middle). The regular integer grid then induces the hexahedral mesh with the desired topology.

Current global parameterizations allow grids to be positioned inside geometrically simple objects whose internal structure (the singularity graph) can be relatively basic. We wish to be able to handle more configurations by improving three aspects of current methods:

- Local grid orientation is usually prescribed by minimizing the curvature of a 3D steering field. Unfortunately, this heuristic does not always provide singularity curves that can be integrated by the parameterization. We plan to explore how to embed integrability constraints in the generation of the direction fields. To address the problem, we already identified necessary validity criteria, for example, the permutation of axes along elementary cycles that go around a singularity must preserve one of the axes (the one tangent to the singularity). The first step to enforce this (necessary) condition will be to split the frame field generation into two parts: first we will define a locally stable vector field, followed by the definition of the other two axes by a 2.5D directional field (2D advected by the stable vector field).
- The grid combinatorial information is characterized by a set of integer coefficients whose values are currently determined through numerical optimization of a geometric criterion: the shape of the hexahedra must be as close as possible to the steering direction field. Thus, the number of layers of hexahedra between two surfaces is determined solely by the size of the hexahedra that one wishes to generate. In this setting degenerate configurations arise easily, and we want to avoid them. In practice, mixed integer solvers often choose to allocate a negative or zero number of layers of hexahedra between two constrained sheets (boundaries of the object, internal constraints or singularities). We will study how to inject strict positivity constraints into these cases, which is a very complex problem because of the subtle interplay between different degrees of freedom of the system. Our first results for quad-meshing of surfaces give promising leads, notably thanks to *motorcycle graphs* [21], a notion we wish to extend to volumes.
- Optimization for the geometric criterion makes it possible to control the average size of the hexahedra, but it does not ensure the bijectivity (even locally) of the resulting parameterizations.

Considering other criteria, as we did in 2D [26], would probably improve the robustness of the process. Our idea is to keep the geometry criterion to find the global topology, but try other criteria to improve the geometry.

3.3. Hexahedral-dominant meshing

All global parameterization approaches are decomposed into three steps: frame field generation, field integration to get a global parameterization, and final mesh extraction. Getting a full hexahedral mesh from a global parameterization means that it has positive Jacobian everywhere except on the frame field singularity graph. To our knowledge, there is no solution to ensure this property, but some efforts are done to limit the proportion of failure cases. An alternative is to produce hexahedral dominant meshes. Our position is in between those two points of view:

1. We want to produce full hexahedral meshes;
2. We consider as pragmatic to keep hexahedral dominant meshes as a fallback solution.

The global parameterization approach yields impressive results on some geometric objects, which is encouraging, but not yet sufficient for numerical analysis. Note that while we attack the remeshing with our parameterizations toolset, the wish to improve the tool itself (as described above) is orthogonal to the effort we put into making the results usable by the industry. To go further, our idea (as opposed to [30], [22]) is that the global parameterization should not handle all the remeshing, but merely act as a guide to fill a large proportion of the domain with a simple structure; it must cooperate with other remeshing bricks, especially if we want to take final application constraints into account.

For each application we will take as an input domains, sets of constraints and, eventually, fields (e.g. the magnetic field in a tokamak). Having established the criteria of mesh quality (per application!) we will incorporate this input into the mesh generation process, and then validate the mesh by a numerical simulation software.

AVIZ Project-Team

3. Research Program

3.1. Scientific Foundations

The scientific foundations of Visual Analytics lie primarily in the domains of Visualization and Data Mining. Indirectly, it inherits from other established domains such as graphic design, Exploratory Data Analysis (EDA), statistics, Artificial Intelligence (AI), Human-Computer Interaction (HCI), and Psychology.

The use of graphic representation to understand abstract data is a goal Visual Analytics shares with Tukey's Exploratory Data Analysis (EDA) [67], graphic designers such as Bertin [54] and Tufte [66], and HCI researchers in the field of Information Visualization [53].

EDA is complementary to classical statistical analysis. Classical statistics starts from a *problem*, gathers *data*, designs a *model* and performs an *analysis* to reach a *conclusion* about whether the data follows the model. While EDA also starts with a problem and data, it is most useful *before* we have a model; rather, we perform visual analysis to discover what kind of model might apply to it. However, statistical validation is not always required with EDA; since often the results of visual analysis are sufficiently clear-cut that statistics are unnecessary.

Visual Analytics relies on a process similar to EDA, but expands its scope to include more sophisticated graphics and areas where considerable automated analysis is required before the visual analysis takes place. This richer data analysis has its roots in the domain of Data Mining, while the advanced graphics and interactive exploration techniques come from the scientific fields of Data Visualization and HCI, as well as the expertise of professions such as cartography and graphic designers who have long worked to create effective methods for graphically conveying information.

The books of the cartographer Bertin and the graphic designer Tufte are full of rules drawn from their experience about how the meaning of data can be best conveyed visually. Their purpose is to find effective visual representation that describe a data set but also (mainly for Bertin) to discover structure in the data by using the right mappings from abstract dimensions in the data to visual ones.

For the last 25 years, the field of Human-Computer Interaction (HCI) has also shown that interacting with visual representations of data in a tight perception-action loop improves the time and level of understanding of data sets. Information Visualization is the branch of HCI that has studied visual representations suitable to understanding and interaction methods suitable to navigating and drilling down on data. The scientific foundations of Information Visualization come from theories about perception, action and interaction.

Several theories of perception are related to information visualization such as the "Gestalt" principles, Gibson's theory of visual perception [59] and Triesman's "preattentive processing" theory [65]. We use them extensively but they only have a limited accuracy for predicting the effectiveness of novel visual representations in interactive settings.

Information Visualization emerged from HCI when researchers realized that interaction greatly enhanced the perception of visual representations.

To be effective, interaction should take place in an interactive loop faster than 100ms. For small data sets, it is not difficult to guarantee that analysis, visualization and interaction steps occur in this time, permitting smooth data analysis and navigation. For larger data sets, more computation should be performed to reduce the data size to a size that may be visualized effectively.

In 2002, we showed that the practical limit of InfoVis was on the order of 1 million items displayed on a screen [57]. Although screen technologies have improved rapidly since then, eventually we will be limited by the physiology of our vision system: about 20 millions receptor cells (rods and cones) on the retina. Another problem will be the limits of human visual attention, as suggested by our 2006 study on change blindness in large and multiple displays [55]. Therefore, visualization alone cannot let us understand very large data sets. Other techniques such as aggregation or sampling must be used to reduce the visual complexity of the data to the scale of human perception.

Abstracting data to reduce its size to what humans can understand is the goal of Data Mining research. It uses data analysis and machine learning techniques. The scientific foundations of these techniques revolve around the idea of finding a good model for the data. Unfortunately, the more sophisticated techniques for finding models are complex, and the algorithms can take a long time to run, making them unsuitable for an interactive environment. Furthermore, some models are too complex for humans to understand; so the results of data mining can be difficult or impossible to understand directly.

Unlike pure Data Mining systems, a Visual Analytics system provides analysis algorithms and processes compatible with human perception and understandable to human cognition. The analysis should provide understandable results quickly, even if they are not ideal. Instead of running to a predefined threshold, algorithms and programs should be designed to allow trading speed for quality and show the tradeoffs interactively. This is not a temporary requirement: it will be with us even when computers are much faster, because good quality algorithms are at least quadratic in time (e.g. hierarchical clustering methods). Visual Analytics systems need different algorithms for different phases of the work that can trade speed for quality in an understandable way.

Designing novel interaction and visualization techniques to explore huge data sets is an important goal and requires solving hard problems, but how can we assess whether or not our techniques and systems provide real improvements? Without this answer, we cannot know if we are heading in the right direction. This is why we have been actively involved in the design of evaluation methods for information visualization [63], [62], [60], [61], [58]. For more complex systems, other methods are required. For these we want to focus on longitudinal evaluation methods while still trying to improve controlled experiments.

3.2. Innovation

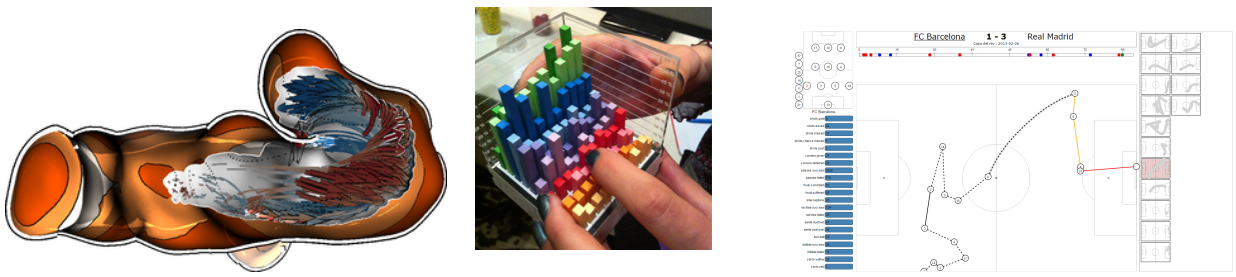


Figure 1. Example novel visualization techniques and tools developed by the team. Left: a non-photorealistic rendering technique that visualizes blood flow and vessel thickness. Middle: a physical visualization showing economic indicators for several countries, right: SoccerStories a tool for visualizing soccer games.

We design novel visualization and interaction techniques (see, for example, Figure 1). Many of these techniques are also evaluated throughout the course of their respective research projects. We cover application domains such as sports analysis, digital humanities, fluid simulations, and biology. A focus of Aviz' work is the improvement of graph visualization and interaction with graphs. We further develop individual techniques

for the design of tabular visualizations and different types of data charts. Another focus is the use of animation as a transition aid between different views of the data. We are also interested in applying techniques from illustrative visualization to visual representations and applications in information visualization as well as scientific visualization [8].

3.3. Evaluation Methods

Evaluation methods are required to assess the effectiveness and usability of visualization and analysis methods. Aviz typically uses traditional HCI evaluation methods, either quantitative (measuring speed and errors) or qualitative (understanding users tasks and activities). Moreover, Aviz is also contributing to the improvement of evaluation methods by reporting on the best practices in the field, by co-organizing workshops (BELIV 2010–2018) to exchange on novel evaluation methods, by improving our ways of reporting, interpreting and communicating statistical results, and by applying novel methodologies, for example to assess visualization literacy [3], [4].

3.4. Software Infrastructures

We want to understand the requirements that software and hardware architectures should provide to support exploratory analysis of large amounts of data. So far, "big data" has been focusing on issues related to storage management and predictive analysis: applying a well-known set of operations on large amounts of data. Visual Analytics is about exploration of data, with sometimes little knowledge of its structure or properties. Therefore, interactive exploration and analysis is needed to build knowledge and apply appropriate analyses; this knowledge and appropriateness is supported by visualizations. However, applying analytical operations on large data implies long-lasting computations, incompatible with interactions, and generates large amounts of results, impossible to visualize directly without aggregation or sampling. Visual Analytics has started to tackle these problems for specific applications but not in a general manner, leading to fragmentation of results and difficulties to reuse techniques from one application to the other. We are interested in abstracting-out the issues and finding general architectural models, patterns, and frameworks to address the Visual Analytics challenge in more generic ways.

3.5. Emerging Technologies

Figure 2. Example emerging technology solutions developed by the team for multi-display environments, wall displays, and token-based visualization.

We want to use different types of display media to empower humans to visually and interactively explore information, in order to better understand and exploit it. This includes novel display equipment and accompanying input techniques. The Aviz team specifically focuses on the exploration of the use of large displays in visualization contexts as well as emerging physical and tangible visualizations (e. g. [6], [5]). In terms of interaction modalities our work focuses on using touch and tangible interaction. Aviz participates to the Digiscope project that funds 11 wall-size displays at multiple places in the Paris area (see http://www.digiscope.fr),

connected by telepresence equipment and a Fablab for creating devices. Aviz is in charge of creating and managing the Fablab, uses it to create physical visualizations, and is also using the local wall-size display (called WILD) to explore visualization on large screens. The team also investigates the perceptual, motor and cognitive implications of using such technologies for visualization.

3.6. Psychology

More cross-fertilization is needed between psychology and information visualization. The only key difference lies in their ultimate objective: understanding the human mind vs. helping to develop better tools. We focus on understanding and using findings from psychology to inform new tools for information visualization. In many cases, our work also extends previous work in psychology. Our approach to the psychology of information visualization is largely holistic and helps bridge gaps between perception, action and cognition in the context of information visualization. Our focus includes the perception of charts in general, perception in large display environments, collaboration, perception of animations, how action can support perception and cognition, and judgment under uncertainty (e. g. [9]).

EX-SITU Project-Team

3. Research Program

3.1. Research Program

We characterize Extreme Situated Interaction as follows:

Extreme users. We study extreme users who make extreme demands on current technology. We know that human beings take advantage of the laws of physics to find creative new uses for physical objects. However, this level of adaptability is severely limited when manipulating digital objects. Even so, we find that creative professionals—artists, designers and scientists—often adapt interactive technology in novel and unexpected ways and find creative solutions. By studying these users, we hope to not only address the specific problems they face, but also to identify the underlying principles that will help us to reinvent virtual tools. We seek to shift the paradigm of interactive software, to establish the laws of interaction that significantly empower users and allow them to control their digital environment.

Extreme situations. We develop extreme environments that push the limits of today's technology. We take as given that future developments will solve "practical" problems such as cost, reliability and performance and concentrate our efforts on interaction in and with such environments. This has been a successful strategy in the past: Personal computers only became prevalent after the invention of the desktop graphical user interface. Smartphones and tablets only became commercially successful after Apple cracked the problem of a usable touch-based interface for the iPhone and the iPad. Although wearable technologies, such as watches and glasses, are finally beginning to take off, we do not believe that they will create the major disruptions already caused by personal computers, smartphones and tablets. Instead, we believe that future disruptive technologies will include fully interactive paper and large interactive displays.

Our extensive experience with the Digiscope WILD and WILDER platforms places us in a unique position to understand the principles of distributed interaction that extreme environments call for. We expect to integrate, at a fundamental level, the collaborative capabilities that such environments afford. Indeed almost all of our activities in both the digital and the physical world take place within a complex web of human relationships. Current systems only support, at best, passive sharing of information, e.g., through the distribution of independent copies. Our goal is to support active collaboration, in which multiple users are actively engaged in the lifecycle of digital artifacts.

Extreme design. We explore novel approaches to the design of interactive systems, with particular emphasis on extreme users in extreme environments. Our goal is to empower creative professionals, allowing them to act as both designers and developers throughout the design process. Extreme design affects every stage, from requirements definition, to early prototyping and design exploration, to implementation, to adaptation and appropriation by end users. We hope to push the limits of participatory design to actively support creativity at all stages of the design lifecycle. Extreme design does not stop with purely digital artifacts. The advent of digital fabrication tools and FabLabs has significantly lowered the cost of making physical objects interactive. Creative professionals now create hybrid interactive objects that can be tuned to the user's needs. Integrating the design of physical objects into the software design process raises new challenges, with new methods and skills to support this form of extreme prototyping.

Our overall approach is to identify a small number of specific projects, organized around four themes: *Creativity, Augmentation, Collaboration* and *Infrastructure*. Specific projects may address multiple themes, and different members of the group work together to advance these different topics.

GRAPHDECO Project-Team

3. Research Program

3.1. Introduction

Our research program is oriented around two main axes: 1) Computer-Assisted Design with Heterogeneous Representations and 2) Graphics with Uncertainty and Heterogeneous Content. These two axes are governed by a set of common fundamental goals, share many common methodological tools and are deeply intertwined in the development of applications.

3.1.1. Computer-Assisted Design with Heterogeneous Representations

Designers use a variety of visual representations to explore and communicate about a concept. Fig. 2 illustrates some typical representations, including sketches, hand-made prototypes, 3D models, 3D printed prototypes or instructions.

(a) Ideation sketch (b) Presentation sketch (c) Coarse prototype (d) 3D model

(f) Simulation (e) 3D Printing (g) Technical diagram (h) Instructions

Figure 2. Various representations of a hair dryer at different stages of the design process. Image source, in order: c-maeng on deviantart.com, shauntur on deviantart.com, "Prototyping and Modelmaking for Product Design" Hallgrimsson, B., Laurence King Publishers, 2012, samsher511 on turbosquid.com, my.solidworks.com, weilung tseng on cargocollective.com, howstuffworks.com, u-manual.com

The early representations of a concept, such as rough sketches and hand-made prototypes, help designers formulate their ideas and test the form and function of multiple design alternatives. These low-fidelity representations are meant to be cheap and fast to produce, to allow quick exploration of the *design space* of the concept. These representations are also often approximate to leave room for subjective interpretation and to stimulate imagination; in this sense, these representations can be considered *uncertain*. As the concept gets more finalized, time and effort are invested in the production of more detailed and accurate representations, such as high-fidelity 3D models suitable for simulation and fabrication. These detailed models can also be used to create didactic instructions for assembly and usage.

Producing these different representations of a concept requires specific skills in sketching, modeling, manufacturing and visual communication. For these reasons, professional studios often employ different experts to produce the different representations of the same concept, at the cost of extensive discussions and numerous iterations between the actors of this process. The complexity of the multi-disciplinary skills involved in the design process also hinders their adoption by laymen.

Existing solutions to facilitate design have focused on a subset of the representations used by designers. However, no solution considers all representations at once, for instance to directly convert a series of sketches into a set of physical prototypes. In addition, all existing methods assume that the concept is unique rather than ambiguous. As a result, rich information about the variability of the concept is lost during each conversion step.

We plan to facilitate design for professionals and laymen by adressing the following objectives:

- We want to assist designers in the exploration of the *design space* that captures the possible variations of a concept. By considering a concept as a *distribution* of shapes and functionalities rather than a single object, our goal is to help designers consider multiple design alternatives more quickly and effectively. Such a representation should also allow designers to preserve multiple alternatives along all steps of the design process rather than committing to a single solution early on and pay the price of this decision for all subsequent steps. We expect that preserving alternatives will facilitate communication with engineers, managers and clients, accelerate design iterations and even allow mass personalization by the end consumers.
- We want to support the various representations used by designers during concept development. While drawings and 3D models have received significant attention in past Computer Graphics research, we will also account for the various forms of rough physical prototypes made to evaluate the shape and functionality of a concept. Depending on the task at hand, our algorithms will either analyse these prototypes to generate a virtual concept, or assist the creation of these prototypes from a virtual model. We also want to develop methods capable of adapting to the different drawing and manufacturing techniques used to create sketches and prototypes. We envision design tools that conform to the habits of users rather than impose specific techniques to them.
- We want to make professional design techniques available to novices. Affordable software, hardware and online instructions are democratizing technology and design, allowing small businesses and individuals to compete with large companies. New manufacturing processes and online interfaces also allow customers to participate in the design of an object via mass personalization. However, similarly to what happened for desktop publishing thirty years ago, desktop manufacturing tools need to be simplified to account for the needs and skills of novice designers. We hope to support this trend by adapting the techniques of professionals and by automating the tasks that require significant expertise.

3.1.2. Graphics with Uncertainty and Heterogeneous Content

Our research is motivated by the observation that traditional CG algorithms have not been designed to account for uncertain data. For example, global illumination rendering assumes accurate virtual models of geometry, light and materials to simulate light transport. While these algorithms produce images of high realism, capturing effects such as shadows, reflections and interreflections, they are not applicable to the growing mass of uncertain data available nowadays.

The need to handle uncertainty in CG is timely and pressing, given the large number of *heterogeneous sources of 3D content* that have become available in recent years. These include data from cheap depth+image sensors (e.g., Kinect or the Tango), 3D reconstructions from image/video data, but also data from large 3D geometry databases, or casual 3D models created using simplified sketch-based modeling tools. Such alternate content has varying levels of *uncertainty* about the scene or objects being modelled. This includes uncertainty in geometry, but also in materials and/or lights – which are often not even available with such content. Since CG algorithms cannot be applied directly, visual effects artists spend hundreds of hours correcting inaccuracies and completing the captured data to make them useable in film and advertising.

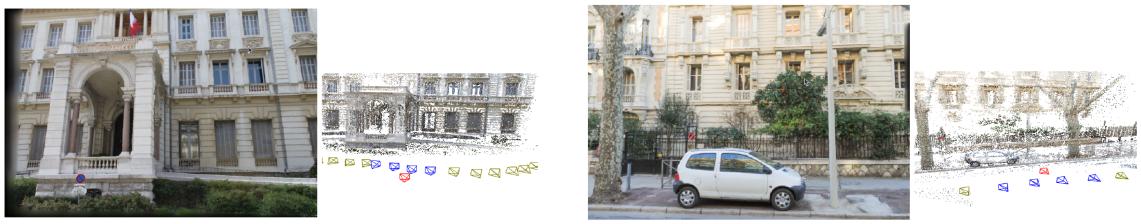


Figure 3. Image-Based Rendering (IBR) techniques use input photographs and approximate 3D to produce new synthetic views.

We identify a major scientific bottleneck which is the need to treat *heterogeneous* content, i.e., containing both (mostly captured) uncertain and perfect, traditional content. Our goal is to provide solutions to this bottleneck, by explicitly and formally modeling uncertainty in CG, and to develop new algorithms that are capable of mixed rendering for this content.

We strive to develop methods in which heterogeneous – and often uncertain – data can be handled automatically in CG with a principled methodology. Our main focus is on *rendering* in CG, including dynamic scenes (video/animations).

Given the above, we need to address the following challenges:

- Develop a theoretical model to handle uncertainty in computer graphics. We must define a new formalism that inherently incorporates uncertainty, and must be able to express traditional CG rendering, both physically accurate and approximate approaches. Most importantly, the new formulation must elegantly handle mixed rendering of perfect synthetic data and captured uncertain content. An important element of this goal is to incorporate *cost* in the choice of algorithm and the optimizations used to obtain results, e.g., preferring solutions which may be slightly less accurate, but cheaper in computation or memory.
- The development of rendering algorithms for heterogeneous content often requires preprocessing of image and video data, which sometimes also includes depth information. An example is the decomposition of images into intrinsic layers of reflectance and lighting, which is required to perform relighting. Such solutions are also useful as image-manipulation or computational photography techniques. The challenge will be to develop such “intermediate” algorithms for the uncertain and heterogeneous data we target.
- Develop efficient rendering algorithms for uncertain and heterogeneous content, reformulating rendering in a probabilistic setting where appropriate. Such methods should allow us to develop approximate rendering algorithms using our formulation in a well-grounded manner. The formalism should include probabilistic models of how the scene, the image and the data interact. These models should be data-driven, e.g., building on the abundance of online geometry and image databases, domain-driven, e.g., based on requirements of the rendering algorithms or perceptually guided, leading to plausible solutions based on limitations of perception.

HYBRID Project-Team

3. Research Program

3.1. Research Program

The scientific objective of Hybrid team is to improve 3D interaction of one or multiple users with virtual environments, by making full use of physical engagement of the body, and by incorporating the mental states by means of brain-computer interfaces. We intend to improve each component of this framework individually, but we also want to improve the subsequent combinations of these components.

The "hybrid" 3D interaction loop between one or multiple users and a virtual environment is depicted in Figure 1 . Different kinds of 3D interaction situations are distinguished (red arrows, bottom): 1) body-based interaction, 2) mind-based interaction, 3) hybrid and/or 4) collaborative interaction (with at least two users). In each case, three scientific challenges arise which correspond to the three successive steps of the 3D interaction loop (blue squares, top): 1) the 3D interaction technique, 2) the modeling and simulation of the 3D scenario, and 3) the design of appropriate sensory feedback.

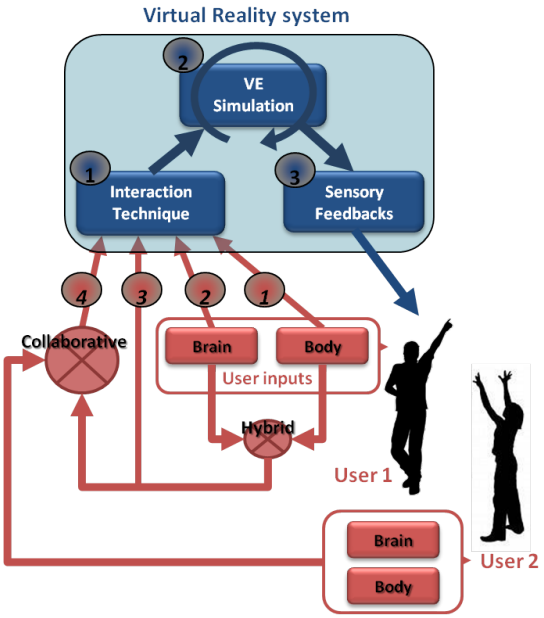


Figure 1. 3D hybrid interaction loop between one or multiple users and a virtual reality system. Top (in blue) three steps of 3D interaction with a virtual environment: (1-blue) interaction technique, (2-blue) simulation of the virtual environment, (3-blue) sensory feedbacks. Bottom (in red) different cases of interaction: (1-red) body-based, (2-red) mind-based, (3-red) hybrid, and (4-red) collaborative 3D interaction.

The 3D interaction loop involves various possible inputs from the user(s) and different kinds of output (or sensory feedback) from the simulated environment. Each user can involve his/her body and mind by means of corporal and/or brain-computer interfaces. A hybrid 3D interaction technique (1) mixes mental and motor inputs and translates them into a command for the virtual environment. The real-time simulation (2) of the

virtual environment is taking into account these commands to change and update the state of the virtual world and virtual objects. The state changes are sent back to the user and perceived by means of different sensory feedbacks (e.g., visual, haptic and/or auditory) (3). The sensory feedbacks are closing the 3D interaction loop. Other users can also interact with the virtual environment using the same procedure, and can eventually "collaborate" by means of "collaborative interactive techniques" (4).

This description is stressing three major challenges which correspond to three mandatory steps when designing 3D interaction with virtual environments:

- **3D interaction techniques:** This first step consists in translating the actions or intentions of the user (inputs) into an explicit command for the virtual environment. In virtual reality, the classical tasks that require such kinds of user command were early categorized in four [44]: navigating the virtual world, selecting a virtual object, manipulating it, or controlling the application (entering text, activating options, etc). The addition of a third dimension, the use of stereoscopic rendering and the use of advanced VR interfaces make however inappropriate many techniques that proved efficient in 2D, and make it necessary to design specific interaction techniques and adapted tools. This challenge is here renewed by the various kinds of 3D interaction which are targeted. In our case, we consider various cases, with motor and/or cerebral inputs, and potentially multiple users.
- **Modeling and simulation of complex 3D scenarios:** This second step corresponds to the update of the state of the virtual environment, in real-time, in response to all the potential commands or actions sent by the user. The complexity of the data and phenomena involved in 3D scenarios is constantly increasing. It corresponds for instance to the multiple states of the entities present in the simulation (rigid, articulated, deformable, fluids, which can constitute both the user's virtual body and the different manipulated objects), and the multiple physical phenomena implied by natural human interactions (squeezing, breaking, melting, etc). The challenge consists here in modeling and simulating these complex 3D scenarios and meeting, at the same time, two strong constraints of virtual reality systems: performance (real-time and interactivity) and genericity (e.g., multi-resolution, multi-modal, multi-platform, etc).
- **Immersive sensory feedbacks:** This third step corresponds to the display of the multiple sensory feedbacks (output) coming from the various VR interfaces. These feedbacks enable the user to perceive the changes occurring in the virtual environment. They are closing the 3D interaction loop, making the user immersed, and potentially generating a subsequent feeling of presence. Among the various VR interfaces which have been developed so far we can stress two kinds of sensory feedback: visual feedback (3D stereoscopic images using projection-based systems such as CAVE systems or Head Mounted Displays); and haptic feedback (related to the sense of touch and to tactile or force-feedback devices). The Hybrid team has a strong expertize in haptic feedback, and in the design of haptic and "pseudo-haptic" rendering [45]. Note that a major trend in the community, which is strongly supported by the Hybrid team, relates to a "perception-based" approach, which aims at designing sensory feedbacks which are well in line with human perceptual capacities.

These three scientific challenges are addressed differently according to the context and the user inputs involved. We propose to consider three different contexts, which correspond to the three different research axes of the Hybrid research team, namely: 1) body-based interaction (motor input only), 2) mind-based interaction (cerebral input only), and then 3) hybrid and collaborative interaction (i.e., the mixing of body and brain inputs from one or multiple users).

3.2. Research Axes

The scientific activity of Hybrid team follows three main axes of research:

- **Body-based interaction in virtual reality.** Our first research axis concerns the design of immersive and effective "body-based" 3D interactions, i.e., relying on a physical engagement of the user's body. This trend is probably the most popular one in VR research at the moment. Most VR setups make use of tracking systems which measure specific positions or actions of the user in order to interact with a virtual environment. However, in recent years, novel options have emerged for measuring

"full-body" movements or other, even less conventional, inputs (e.g. body equilibrium). In this first research axis we are thus concerned by the emergence of new kinds of "body-based interaction" with virtual environments. This implies the design of novel 3D user interfaces and novel 3D interactive techniques, novel simulation models and techniques, and novel sensory feedbacks for body-based interaction with virtual worlds. It involves real-time physical simulation of complex interactive phenomena, and the design of corresponding haptic and pseudo-haptic feedback.

- **Mind-based interaction in virtual reality.** Our second research axis concerns the design of immersive and effective "mind-based" 3D interactions in Virtual Reality. Mind-based interaction with virtual environments is making use of Brain-Computer Interface technology. This technology corresponds to the direct use of brain signals to send "mental commands" to an automated system such as a robot, a prosthesis, or a virtual environment. BCI is a rapidly growing area of research and several impressive prototypes are already available. However, the emergence of such a novel user input is also calling for novel and dedicated 3D user interfaces. This implies to study the extension of the mental vocabulary available for 3D interaction with VE, then the design of specific 3D interaction techniques "driven by the mind" and, last, the design of immersive sensory feedbacks that could help improving the learning of brain control in VR.
- **Hybrid and collaborative 3D interaction.** Our third research axis intends to study the combination of motor and mental inputs in VR, for one or multiple users. This concerns the design of mixed systems, with potentially collaborative scenarios involving multiple users, and thus, multiple bodies and multiple brains sharing the same VE. This research axis therefore involves two interdependent topics: 1) collaborative virtual environments, and 2) hybrid interaction. It should end up with collaborative virtual environments with multiple users, and shared systems with body and mind inputs.

ILDA Project-Team

3. Research Program

3.1. Introduction

Our ability to acquire or generate, store, process, interlink and query data has increased spectacularly over the last few years. The corresponding advances are commonly grouped under the umbrella of so called *Big Data*. Even if the latter has become a buzzword, these advances are real, and they are having a profound impact in domains as varied as scientific research, commerce, social media, industrial processes or e-government. Yet, looking ahead, emerging technologies related to what we now call the *Web of Data* (a.k.a the Semantic Web) have the potential to create an even larger revolution in data-driven activities, by making information accessible to machines as semistructured data [26] that eventually becomes actionable knowledge. Indeed, novel Web data models considerably ease the interlinking of semi-structured data originating from multiple independent sources. They make it possible to associate machine-processable semantics with the data. This in turn means that heterogeneous systems can exchange data, infer new data using reasoning engines, and that software agents can cross data sources, resolving ambiguities and conflicts between them [77]. Datasets are becoming very rich and very large. They are gradually being made even larger and more heterogeneous, but also much more useful, by interlinking them, as exemplified by the Linked Data initiative [49].

These advances raise research questions and technological challenges that span numerous fields of computer science research: databases, communication networks, security and trust, data mining, as well as human-computer interaction. Our research is based on the conviction that interactive systems play a central role in many data-driven activity domains. Indeed, no matter how elaborate the data acquisition, processing and storage pipelines are, data eventually get processed or consumed one way or another by users. The latter are faced with large, increasingly interlinked heterogeneous datasets (see, *e.g.*, Figure 1) that are organized according to complex structures, resulting in overwhelming amounts of both raw data and structured information. Users thus require effective tools to make sense of their data and manipulate them.

Figure 1. Linking Open Data cloud diagram from 2007 to 2017 – http://lod-cloud.net

We approach this problem from the perspective of the Human-Computer Interaction (HCI) field of research, whose goal is to study how humans interact with computers and inspire novel hardware and software designs aimed at optimizing properties such as efficiency, ease of use and learnability, in single-user or cooperative work contexts. More formally, HCI is about designing systems that lower the barrier between users' cognitive model of what they want to accomplish, and computers' understanding of this model. HCI is about the design, implementation and evaluation of computing systems that humans interact with [54], [79]. It is a highly multidisciplinary field, with experts from computer science, cognitive psychology, design, engineering, ethnography, human factors and sociology.

In this broad context, ILDA aims at designing interactive systems that display [35], [61], [87] the data and let users interact with them, aiming to help users better *navigate* and *comprehend* large webs of data represented visually [7], as well as *relate* and *manipulate* them.

Our research agenda consists of the three complementary axes detailed in the following subsections. Designing systems that consider interaction in close conjunction with data semantics is pivotal to all three axes. Those semantics will help drive navigation in, and manipulation of, the data, so as to optimize the communication bandwidth between users and data.

3.2. Semantics-driven Data Manipulation

Participants: Emmanuel Pietriga, Caroline Appert, Anastasia Bezerianos, Marie Destandau, Hugo Romat, Tong Xue, Léo Colombaro.

The Web of Data has been maturing for the last fifteen years and is starting to gain adoption across numerous application domains (Figure 1). Now that most foundational building blocks are in place, from knowledge representation, inference mechanisms and query languages [50], all the way up to the expression of data presentation knowledge [70] and to mechanisms like look-up services [86] or spreading activation [43], we need to pay significant attention to how human beings are going to interact with this new Web, if it is to *"reach its full potential"* [44].

Most efforts in terms of user interface design and development for the Web of data have essentially focused on tools for software developers or subject-matter experts who create ontologies and populate them [56], [41]. Tools more oriented towards end-users are starting to appear [32], [34], [51], [52], [55], [64], including the so-called *linked data browsers* [49]. However, those browsers are in most cases based on quite conventional point-and-click hypertext interfaces that present data to users in a very page-centric, web-of-documents manner that is ill-suited to navigating in, and manipulating, webs of data.

To be successful, interaction paradigms that let users navigate and manipulate data on the Web have to be tailored to the radically different way of browsing information enabled by it, where users directly interact with the data rather than with monolithic documents. The general research question addressed in this part of our research program is how to design novel interaction techniques that help users manipulate their data more efficiently. By data manipulation, we mean all low-level tasks related to manually creating new content, modifying and cleaning existing content, merging data from different sources, establishing connections between datasets, categorizing data, and eventually sharing the end results with other users; tasks that are currently considered quite tedious because of the sheer complexity of the concepts, data models and syntax, and the interplay between all of them.

Our approach is based on the conviction that there is a strong potential for cross-fertilization, as mentioned earlier: on the one hand, user interface design is essential to the management and understanding of webs of data; on the other hand, interlinked datasets enriched with even a small amount of semantics can help create more powerful user interfaces, that provide users with the right information at the right time.

We envision systems that focus on the data themselves, exploiting the underlying *semantics and structure* in the background rather than exposing them – which is what current user interfaces for the Web of Data often do. We envision interactive systems in which the semantics and structure are not exposed directly to users, but serve as input to the system to generate interactive representations that convey information relevant to the task at hand and best afford the possible manipulation actions.

Relevant publications by team members this year: [22], [15], [17] and major ones in recent years: [7].

3.3. Generalized Multi-scale Navigation

Participants: Caroline Appert, Anastasia Bezerianos, Olivier Chapuis, Emmanuel Pietriga, Vanessa Peña-Araya, Marie Destandau, Anna Gogolou, Hugo Romat, Dylan Lebout.

The foundational question addressed here is what to display when, where and how, so as to provide effective support to users in their data understanding and manipulation tasks. ILDA targets contexts in which workers have to interact with complementary views on the same data, or with views on different-but-related datasets, possibly at different levels of abstraction. Being able to combine or switch between representations of the data at different levels of detail and merge data from multiple sources in a single representation is central to many scenarios. This is especially true in both of the application domains we consider: mission-critical systems (e.g., natural disaster crisis management) and the exploratory analysis of scientific data (e.g., correlate theories and heterogeneous observational data for an analysis of a given celestial body in Astrophysics).

A significant part of our research over the last ten years has focused on multi-scale interfaces. We designed and evaluated novel interaction techniques, but also worked actively on the development of open-source UI toolkits for multi-scale interfaces (http://zvtm.sf.net). These interfaces let users navigate large but relatively homogeneous datasets at different levels of detail, on both workstations [73], [29], [69], [68], [67], [30], [72], [28], [74] and wall-sized displays [63], [58], [71], [62], [31], [37], [36]. This part of the ILDA research program is about extending multi-scale navigation in two directions: 1. Enabling the representation of multiple, spatially-registered but widely varying, multi-scale data layers in Geographical Information Systems (GIS); 2. Generalizing the multi-scale navigation paradigm to interconnected, heterogeneous datasets as found on the Web of Data.

The first research problem has been mainly investigated in collaboration with IGN in the context of ANR project MapMuxing, which stands for *multi-dimensional map multiplexing*, from 2014 to early 2019. Project MapMuxing aimed at going beyond the traditional pan & zoom and overview+detail interface schemes, and at designing and evaluating novel cartographic visualizations that rely on high-quality generalization, *i.e.*, the simplification of geographic data to make it legible at a given map scale [82], [83], and symbol specification. Beyond project MapMuxing, we are also investigating multi-scale multiplexing techniques for geo-localized data in the specific context of ultra-high-resolution wall-sized displays, where the combination of a very high pixel density and large physical surface enable us to explore designs that involve collaborative interaction and physical navigation in front of the workspace. This is work done in cooperation with team Massive Data at Inria Chile.

The second research problem is about the extension of multi-scale navigation to interconnected, heterogeneous datasets. Generalization has a rather straightforward definition in the specific domain of geographical information systems, where data items are geographical entities that naturally aggregate as scale increases. But it is unclear how generalization could work for representations of the more heterogeneous webs of data that we consider in the first axis of our research program. Those data form complex networks of resources with multiple and quite varied relationships between them, that cannot rely on a single, unified type of representation (a role played by maps in GIS applications).

Addressing the limits of current generalization processes is a longer-term, more exploratory endeavor. Here again, the machine-processable semantics and structure of the data give us an opportunity to rethink how users navigate interconnected heterogeneous datasets. Using these additional data, we investigate ways to generalize the multi-scale navigation paradigm to datasets whose layout and spatial relationships can be much richer and much more diverse than what can be encoded with static linear hierarchies as typically found today in interfaces for browsing maps or large imagery. Our goal is thus to design and develop highly dynamic and versatile multi-scale information spaces for heterogeneous data whose structure and semantics are not known in advance, but discovered incrementally.

Relevant publications by team members this year: [24], [20], [13], [14], [11], [19] and major ones in recent years: [10], [2].

3.4. Novel Forms of Input for Groups and Individuals

Participants: Caroline Appert, Anastasia Bezerianos, Olivier Chapuis, Emmanuel Pietriga, Eugénie Brasier, Emmanuel Courtoux, Raphaël James.

Analyzing and manipulating large datasets can involve multiple users working together in a coordinated manner in multi-display environments: workstations, handheld devices, wall-sized displays [31]. Those users work towards a common goal, navigating and manipulating data displayed on various hardware surfaces in a coordinated manner. Group awareness [48], [25] is central in these situations, as users, who may or may not be co-located in the same room, can have an optimal individual behavior only if they have a clear picture of what their collaborators have done and are currently doing in the global context. We work on the design and implementation of interactive systems that improve group awareness in co-located situations [57], making individual users able to figure out what other users are doing without breaking the flow of their own actions.

In addition, users need a rich interaction vocabulary to handle large, structured datasets in a flexible and powerful way, regardless of the context of work. Input devices such as mice and trackpads provide a limited number of input actions, thus requiring users to switch between modes to perform different types of data manipulation and navigation actions. The action semantics of these input devices are also often too much dependent on the display output. For instance, a mouse movement and click can only be interpreted according to the graphical controller (widget) above which it is moved. We focus on designing powerful input techniques based upon technologies such as tactile surfaces (supported by UI toolkits developed in-house), 3D motion tracking systems, or custom-built controllers [60] *to complement (rather than replace) traditional input devices* such as keyboards, that remain the best method so far for text entry, and indirect input devices such as mice or trackpads for pixel-precise pointing actions.

The input vocabularies we investigate enable users to navigate and manipulate large and structured datasets in environments that involve multiple users and displays that vary in their size, position and orientation [31], [45], each having their own characteristics and affordances: wall displays [63], [89], workstations, tabletops [66], [40], tablets [65], [84], smartphones [88], [38], [80], [81], and combinations thereof [39], [85], [62], [31].

We aim at designing rich interaction vocabularies that go far beyond what current touch interfaces offer, which rarely exceeds five gestures such as simple slides and pinches. Designing larger gesture vocabularies requires identifying discriminating dimensions (e.g., the presence or absence of anchor points and the distinction between internal and external frames of reference [65]) in order to structure a space of gestures that interface designers can use as a dictionary for choosing a coherent set of controls. These dimensions should be few and simple, so as to provide users with gestures that are easy to memorize and execute. Beyond gesture complexity, the scalability of vocabularies also depends on our ability to design robust gesture recognizers that will allow users to fluidly chain simple gestures that make it possible to interlace navigation and manipulation actions.

We also study how to further extend input vocabularies by combining touch [65], [88], [66] and mid-air gestures [63] with physical objects [53], [78], [60] and classical input devices such as keyboards to enable users to input commands to the system or to involve other users in their workflow (request for help, delegation, communication of personal findings, etc.) [33], [59]. Gestures and objects encode a lot of information in their shape, dynamics and direction, that can be directly interpreted in relation with the user, independently from the display output. Physical objects can also greatly improve coordination among actors for, e.g., handling priorities or assigning specific roles.

Relevant publications by team members this year: [9], [23], [16], [22] and major ones in recent years: [1], [10], [5], [3], [8].

IMAGINE Project-Team

3. Research Program

3.1. Methodology

As already stressed, thinking of future digital modeling technologies as an Expressive Virtual Pen enabling to seamlessly design, refine and convey animated 3D content, leads to revisit models for shapes, motions and stories from a user-centered perspective. More specifically, inspiring from the user-centered interfaces developed in the Human Computer Interaction domain, we introduced the new concept of user-centered graphical models. Ideally, such models should be designed to behave, under any user action, the way a human user would have predicted. In our case, user's actions may include creation gestures such as sketching to draft a shape or direct a motion, deformation gestures such as stretching a shape in space or a motion in time, or copy-paste gestures to transfer some of the features from existing models to other ones. User-centered graphical models need to incorporate knowledge in order to seamlessly generate the appropriate content from such actions. We are using the following methodology to advance towards these goals:

- Develop high-level models for shapes, motion and stories that embed the necessary knowledge to respond as expected to user actions. These models should provide the appropriate handles for conveying the user's intent while embedding procedural methods that seamlessly take care of the appropriate details and constraints.
- Combine these models with expressive design and control tools such as gesture-based control through sketching, sculpting, or acting, towards interactive environments where users can create a new virtual scene, play with it, edit or refine it, and semi-automatically convey it through a video.

3.2. Validation

Validation is a major challenge when developing digital creation tools: there is no ideal result to compare with, in contrast with more standard problems such as reconstructing existing shapes or motions. Therefore, we had to think ahead about our validation strategy: new models for geometry or animation can be validated, as usually done in Computer Graphics, by showing that they solve a problem never tackled before or that they provide a more general or more efficient solution than previous methods. The interaction methods we are developing for content creation and editing rely as much as possible on existing interaction design principles already validated within the HCI community. We also occasionally develop new interaction tools, most often in collaboration with this community, and validate them through user studies. Lastly, we work with expert users from various application domains through our collaborations with professional artists, scientists from other domains, and industrial partners: these expert users validate the use of our new tools compared to their usual pipeline.

LOKI Project-Team

3. Research Program

3.1. Introduction

Interaction is by nature a dynamic phenomenon that takes place between interactive systems and their users. Redesigning interactive systems to better account for interaction requires fine understanding of these dynamics from the user side so as to better handle them from the system side. In fact, layers of actual interactive systems abstract hardware and system resources from a system and programing perspective. Following our Interaction Machine concept, we are reconsidering these architectures from the user's perspective, through different *levels of dynamics of interaction* (see Figure 1).

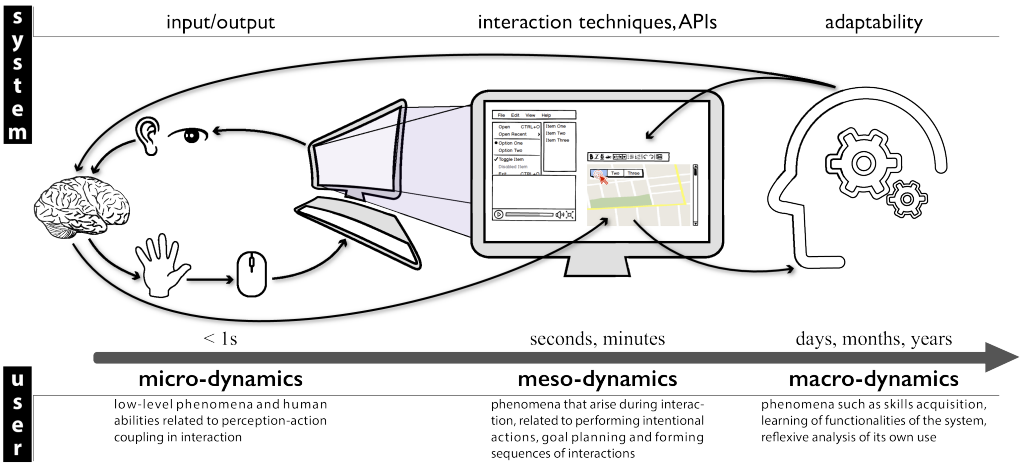


Figure 1. Levels of dynamics of interaction.

Considering phenomena that occur at each of these levels as well as their relationships will help us to acquire the necessary knowledge (Empowering Tools) and technological bricks (Interaction Machine) to reconcile the way interactive systems are designed and engineered with human abilities. Although our strategy is to investigate issues and address challenges for all of the three levels, our immediate priority is to focus on micro-dynamics since it concerns very fundamental knowledge about interaction and relates to very low-level parts of interactive systems, which is likely to influence our future research and developments at the other levels.

3.2. Micro-Dynamics

Micro-dynamics involve low-level phenomena and human abilities which are related to short time/instantness and to perception-action coupling in interaction, when the user has almost no control or consciousness of the action once it has been started. From a system perspective, it has implications mostly on input and output (I/O) management.

3.2.1. Transfer functions design and latency management

We have developed a recognized expertise in the characterization and the design of *transfer functions* [34], [45], i. e., the algorithmic transformations of raw user input for system use. Ideally, transfer functions should match the interaction context. Yet the question of how to maximize one or more criteria in a given context remains an open one, and on-demand adaptation is difficult because transfer functions are usually implemented at the lowest possible level to avoid latency. Latency has indeed long been known as a determinant of human performance in interactive systems [41] and recently regained attention with touch interactions [40]. These two problems require cross examination to improve performance with interactive systems: Latency can be a confounding factor when evaluating the effectiveness of transfer functions, and transfer functions can also include algorithms to compensate for latency.

We have recently proposed new cheap but robust methods for the measurement of end-to-end latency [2] and are currently working on compensation methods and the evaluation of their perceived side effects. Our goal is then to automatically adapt the transfer function to individual users and contexts of use while reducing latency in order to support stable and appropriate control. To achieve this, we will investigate combinations of low-level (embedded) and high-level (application) ways to take user capabilities and task characteristics into account and reduce or compensate for latency in different contexts, e. g., using a mouse or a touchpad, a touch-screen, an optical finger navigation device or a brain-computer interface. From an engineering perspective, this knowledge on low-level human factors will help us to rethink and redesign the I/O loop of interactive systems in order to better account for them and achieve more adapted and adaptable perception-action coupling.

3.2.2. Tactile feedback & haptic perception

We are also concerned with the physicality of human-computer interaction, with a focus on haptic perception and related technologies. For instance, when interacting with virtual objects such as software buttons on a touch surface, the user cannot feel the click sensation as with physical buttons. The tight coupling between how we perceive and how we manipulate objects is then essentially broken although this is instrumental for efficient direct manipulation. We have addressed this issue in multiple contexts by designing, implementing and evaluating novel applications of tactile feedback [5].

In comparison with many other modalities, one difficulty with tactile feedback is its diversity. It groups sensations of forces, vibrations, friction, or deformation. Although this is a richness, it also raises usability and technological challenges since each kind of haptic stimulation requires different kinds of actuators with their own parameters and thresholds. And results from one are hardly applicable to others. On a “knowledge” point of view, we want to better understand and empirically classify haptic variables and the kind of information they can represent (continuous, ordinal, nominal), their resolution, and their applicability to various contexts. From the “technology” perspective, we want to develop tools to inform and ease the design of haptic interactions taking best advantage of the different technologies in a consistent and transparent way.

3.3. Meso-Dynamics

Meso-dynamics relate to phenomena that arise during interaction, on a longer but still short time-scale. For users, it is related to performing intentional actions, to goal planning and tools selection, and to forming sequences of interactions based on a known set of rules or instructions. From the system perspective, it relates to how possible actions are exposed to the user and how they have to be executed (i. e., interaction techniques). It also has implication on the tools for designing and implementing those techniques (programming languages and APIs).

3.3.1. Interaction bandwidth and vocabulary

Interactive systems and their applications have an always-increasing number of available features and commands due to, e. g., the large amount of data to manipulate, increasing power and number of functionalities, or multiple contexts of use.

On the input side, we want to augment the *interaction bandwidth* between the user and the system in order to cope with this increasing complexity. In fact, most input devices capture only a few of the movements and actions the human body is capable of. Our arms and hands for instance have many degrees of freedom that are not fully exploited in common interfaces. We have recently designed new technologies to improve expressibility such as a bendable digitizer pen [36], or reliable technology for studying the benefits of finger identification on multi-touch interfaces [4].

On the output side, we want to expand users' *interaction vocabulary*. All of the features and commands of a system can not be displayed on screen at the same time and lots of *advanced* features are by default hidden to the users (e. g., hotkeys) or buried in deep hierarchies of command-triggering systems (e. g., menus). As a result, users tend to use only a subset of all the tools the system actually offers [44]. We will study how to help them to broaden their knowledge of available functions.

Through this "opportunistic" exploration of alternative and more expressive input methods and interaction techniques, we will particularly focus on the necessary technological requirements to integrate them into interactive systems, in relation with our redesign of the I/O stack at the micro-dynamics level.

3.3.2. Spatial and temporal continuity in interaction

At a higher-level, we will investigate how more expressive interaction techniques affect users' strategies when performing sequences of elementary actions and tasks. More generally, we will explore the "*continuity*" in interaction. Interactive systems have moved from one computer to multiple connected interactive devices (computer, tablets, phones, watches, etc.) that could also be augmented through a Mixed-Reality paradigm. This distribution of interaction raises new challenges from both usability and engineering perspectives that we clearly have to consider in our main objective of revisiting interactive systems [43]. It involves the simultaneous use of multiple devices and also the changes in the role of devices according to the location, the time, the task, and contexts of use: a tablet device can be used as the main device while traveling, and it becomes an input device or a secondary monitor for continuing the same task once in the office; a smart-watch can be used as a standalone device to send messages, but also as a remote controller for a wall-sized display. One challenge is then to design interaction techniques that support seamless and smooth transitions during these spatial and temporal changes of the system in order to maintain the continuity of uses and tasks, and how to integrate these principles in future interactive systems.

3.3.3. Expressive tools for prototyping, studying, and programming interaction

Current systems suffer from engineering issues that keep constraining and influencing how interaction is thought, designed, and implemented. Addressing the challenges we presented in this section and making the solutions possible require extended expressiveness, and researchers and designers must either wait for the proper toolkits to appear, or "hack" existing interaction frameworks, often bypassing existing mechanisms. For instance, numerous usability problems in existing interfaces stem from a common cause: the lack, or untimely discarding, of relevant information about how events are propagated and how changes come to occur in interactive environments. On top of our redesign of the I/O loop of interactive systems, we will investigate how to facilitate access to that information and also promote a more grounded and expressive way to describe and exploit input-to-output chains of events at every system level. We want to provide finer granularity and better-described connections between the *causes* of changes (e.g. input events and system triggers), their *context* (e.g. system and application states), their *consequences* (e.g. interface and data updates), and their *timing* [8]. More generally, a central theme of our Interaction Machine vision is to promote interaction as a first-class object of the system [33], and we will study alternative and better-adapted technologies for designing and programming interaction, such as we did recently to ease the prototyping of Digital Musical Instruments [1] or the programming of animations in graphical interfaces [10]. Ultimately, we want to propose a unified model of hardware and software scaffolding for interaction that will contribute to the design of our Interaction Machine.

3.4. Macro-Dynamics

Macro-dynamics involve longer-term phenomena such as skills acquisition, learning of functionalities of the system, reflexive analysis of its own use (e. g., when the user has to face novel or unexpected situations which require high-level of knowledge of the system and its functioning). From the system perspective, it implies to better support cross-application and cross-platform mechanisms so as to favor skill transfer. It also requires to improve the instrumentation and high-level logging capabilities to favor reflexive use, as well as flexibility and adaptability for users to be able to finely tune and shape their tools.

We want to move away from the usual binary distinction between "novices" and "experts" [3] and explore means to promote and assist digital skill acquisition in a more progressive fashion. Indeed, users have a permanent need to adapt their skills to the constant and rapid evolution of the tasks and activities they carry on a computer system, but also the changes in the software tools they use [47]. Software strikingly lacks powerful means of acquiring and developing these skills [3], forcing users to mostly rely on outside support (e. g., being guided by a knowledgeable person, following online tutorials of varying quality). As a result, users tend to rely on a surprisingly limited interaction vocabulary, or *make-do* with sub-optimal routines and tools [48]. Ultimately, the user should be able to master the interactive system to form durable and stabilized practices that would eventually become *automatic* and reduce the mental and physical efforts, making their interaction *transparent*.

In our previous work, we identified the fundamental factors influencing expertise development in graphical user interfaces, and created a conceptual framework that characterizes users' performance improvement with UIs [7], [3]. We designed and evaluated new command selection and learning methods to leverage user's digital skill development with user interfaces, on both desktop [6] and touch-based computers.

We are now interested in broader means to support the analytic use of computing tools:

- *to foster understanding of interactive systems.* As the digital world makes the shift to more and more complex systems driven by machine learning algorithms, we increasingly lose our comprehension of which process caused the system to respond in one way rather than another. We will study how novel interactive visualizations can help reveal and expose the "intelligence" behind, in ways that people better master their complexity.
- *to foster reflexion on interaction.* We will study how we can foster users' reflexion on their own interaction in order to encourage them to acquire novel digital skills. We will build real-time and off-line software for monitoring how user's ongoing activity is conducted at an application and system level. We will develop augmented feedbacks and interactive history visualization tools that will offer contextual visualizations to help users to better understand and share their activity, compare their actions to that of others, and discover possible improvement.
- *to optimize skill-transfer and tool re-appropriation.* The rapid evolution of new technologies has drastically increased the frequency at which systems are updated, often requiring to relearn everything from scratch. We will explore how we can minimize the cost of having to appropriate an interactive tool by helping users to capitalize on their existing skills.

We plan to explore these questions as well as the use of such aids in several contexts like web-based, mobile, or BCI-based applications. Although, a core aspect of this work will be to design systems and interaction techniques that will be as little platform-specific as possible, in order to better support skill transfer. Following our Interaction Machine vision, this will lead us to rethink how interactive systems have to be engineered so that they can offer better instrumentation, higher adaptability, and fewer separation between applications and tasks in order to support reuse and skill transfer.

MANAO Project-Team

3. Research Program

3.1. Related Scientific Domains

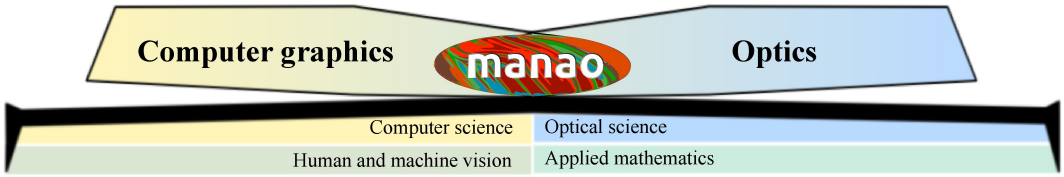


Figure 3. Related scientific domains of the MANAO project.

The *MANAO* project aims at studying, acquiring, modeling, and rendering the interactions between the three components that are light, shape, and matter from the viewpoint of an observer. As detailed more lengthily in the next section, such a work will be done using the following approach: first, we will tend to consider that these three components do not have strict frontiers when considering their impacts on the final observers; then, we will not only work in **computer graphics**, but also at the intersection of computer graphics and **optics**, exploring the mutual benefits that the two domains may provide. It is thus intrinsically a **transdisciplinary** project (as illustrated in Figure 3) and we expect results in both domains.

Thus, the proposed team-project aims at establishing a close collaboration between computer graphics (e.g., 3D modeling, geometry processing, shading techniques, vector graphics, and GPU programming) and optics (e.g., design of optical instruments, and theories of light propagation). The following examples illustrate the strengths of such a partnership. First, in addition to simpler radiative transfer equations [36] commonly used in computer graphics, research in the later will be based on state-of-the-art understanding of light propagation and scattering in real environments. Furthermore, research will rely on appropriate instrumentation expertise for the measurement [48], [49] and display [47] of the different phenomena. Reciprocally, optics researches may benefit from the expertise of computer graphics scientists on efficient processing to investigate interactive simulation, visualization, and design. Furthermore, new systems may be developed by unifying optical and digital processing capabilities. Currently, the scientific background of most of the team members is related to computer graphics and computer vision. A large part of their work have been focused on simulating and analyzing optical phenomena as well as in acquiring and visualizing them. Combined with the close collaboration with the optics laboratory LP2N (http://www.lp2n.fr) and with the students issued from the "Institut d'Optique" (http://www.institutoptique.fr), this background ensures that we can expect the following results from the project: the construction of a common vocabulary for tightening the collaboration between the two scientific domains and creating new research topics. By creating this context, we expect to attract (and even train) more trans-disciplinary researchers.

At the boundaries of the *MANAO* project lie issues in **human and machine vision**. We have to deal with the former whenever a human observer is taken into account. On one side, computational models of human vision are likely to guide the design of our algorithms. On the other side, the study of interactions between light, shape, and matter may shed some light on the understanding of visual perception. The same kind of connections are expected with machine vision. On the one hand, traditional computational methods for acquisition (such as photogrammetry) are going to be part of our toolbox. On the other hand, new display technologies (such as the ones used for augmented reality) are likely to benefit from our integrated approach

and systems. In the *MANAO* project we are mostly users of results from human vision. When required, some experimentation might be done in collaboration with experts from this domain, like with the European PRISM project. For machine vision, provided the tight collaboration between optical and digital systems, research will be carried out inside the *MANAO* project.

Analysis and modeling rely on **tools from applied mathematics** such as differential and projective geometry, multi-scale models, frequency analysis [38] or differential analysis [70], linear and non-linear approximation techniques, stochastic and deterministic integrations, and linear algebra. We not only rely on classical tools, but also investigate and adapt recent techniques (e.g., improvements in approximation techniques), focusing on their ability to run on modern hardware: the development of our own tools (such as Eigen) is essential to control their performances and their abilities to be integrated into real-time solutions or into new instruments.

3.2. Research axes

The *MANAO* project is organized around four research axes that cover the large range of expertise of its members and associated members. We briefly introduce these four axes in this section. More details and their inter-influences that are illustrated in the Figure 2 will be given in the following sections.

Axis 1 is the theoretical foundation of the project. Its main goal is to increase the understanding of light, shape, and matter interactions by combining expertise from different domains: optics and human/machine vision for the analysis and computer graphics for the simulation aspect. The goal of our analyses is to identify the different layers/phenomena that compose the observed signal. In a second step, the development of physical simulations and numerical models of these identified phenomena is a way to validate the pertinence of the proposed decompositions.

In Axis 2, the final observers are mainly physical captors. Our goal is thus the development of new acquisition and display technologies that combine optical and digital processes in order to reach fast transfers between real and digital worlds, in order to increase the convergence of these two worlds.

Axes 3 and 4 focus on two aspects of computer graphics: rendering, visualization and illustration in Axis 3, and editing and modeling (content creation) in Axis 4. In these two axes, the final observers are mainly human users, either generic users or expert ones (e.g., archaeologist [74], computer graphics artists).

3.3. Axis 1: Analysis and Simulation

Challenge: Definition and understanding of phenomena resulting from interactions between light, shape, and matter as seen from an observer point of view.

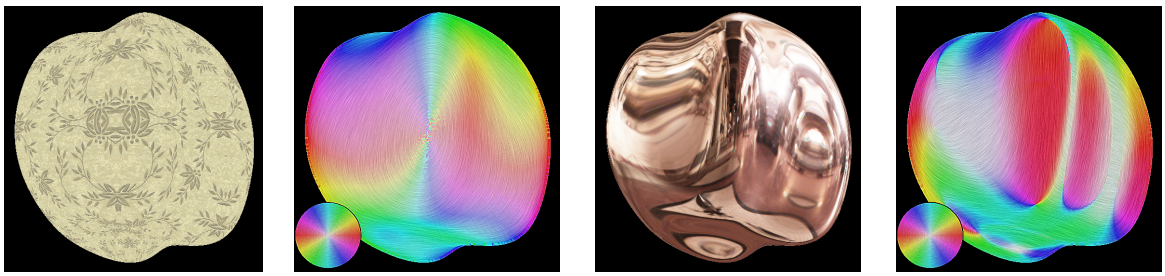
Results: Theoretical tools and numerical models for analyzing and simulating the observed optical phenomena.

To reach the goals of the *MANAO* project, we need to **increase our understanding** of how light, shape, and matter act together in synergy and how the resulting signal is finally observed. For this purpose, we need to identify the different phenomena that may be captured by the targeted observers. This is the main objective of this research axis, and it is achieved by using three approaches: the simulation of interactions between light, shape, and matter, their analysis and the development of new numerical models. This resulting improved knowledge is a foundation for the researches done in the three other axes, and the simulation tools together with the numerical models serve the development of the joint optical/digital systems in Axis 2 and their validation.

One of the main and earliest goals in computer graphics is to faithfully reproduce the real world, focusing mainly on light transport. Compared to researchers in physics, researchers in computer graphics rely on a subset of physical laws (mostly radiative transfer and geometric optics), and their main concern is to efficiently use the limited available computational resources while developing as fast as possible algorithms. For this purpose, a large set of theoretical as well as computational tools has been introduced to take a **maximum benefit of hardware** specificities. These tools are often dedicated to specific phenomena (e.g., direct or indirect lighting, color bleeding, shadows, caustics). An efficiency-driven approach needs such a classification of light paths [44] in order to develop tailored strategies [86]. For instance, starting from simple direct lighting,

more complex phenomena have been progressively introduced: first diffuse indirect illumination [42], [78], then more generic inter-reflections [51], [36] and volumetric scattering [75], [33]. Thanks to this search for efficiency and this classification, researchers in computer graphics have developed a now recognized expertise in fast-simulation of light propagation. Based on finite elements (radiosity techniques) or on unbiased Monte Carlo integration schemes (ray-tracing, particle-tracing, ...), the resulting algorithms and their combination are now sufficiently accurate to be used-back in physical simulations. The *MANAO* project will continue the search for **efficient and accurate simulation** techniques, but extending it from computer graphics to optics. Thanks to the close collaboration with scientific researchers from optics, new phenomena beyond radiative transfer and geometric optics will be explored.

Search for algorithmic efficiency and accuracy has to be done in parallel with **numerical models**. The goal of visual fidelity (generalized to accuracy from an observer point of view in the project) combined with the goal of efficiency leads to the development of alternative representations. For instance, common classical finite-element techniques compute only basis coefficients for each discretization element: the required discretization density would be too large and to computationally expensive to obtain detailed spatial variations and thus visual fidelity. Examples includes texture for decorrelating surface details from surface geometry and high-order wavelets for a multi-scale representation of lighting [32]. The numerical complexity explodes when considering directional properties of light transport such as radiance intensity (Watt per square meter and per steradian - $W.m^{-2}.sr^{-1}$), reducing the possibility to simulate or accurately represent some optical phenomena. For instance, Haar wavelets have been extended to the spherical domain [77] but are difficult to extend to non-piecewise-constant data [80]. More recently, researches prefer the use of Spherical Radial Basis Functions [83] or Spherical Harmonics [69]. For more complex data, such as reflective properties (e.g., BRDF [63], [52] - 4D), ray-space (e.g., Light-Field [60] - 4D), spatially varying reflective properties (6D - [73]), new models, and representations are still investigated such as rational functions [66] or dedicated models [20] and parameterizations [76], [81]. For each (newly) defined phenomena, we thus explore the space of possible numerical representations to determine the **most suited one for a given application**, like we have done for BRDF [66].



Texuring | 1st order gradient field | Environment reflection | 2st order gradient field

Figure 4. First-oder analysis [87] have shown that shading variations are caused by depth variations (first-order gradient field) and by normal variations (second-order fields). These fields are visualized using hue and saturation to indicate direction and magnitude of the flow respectively.

Before being able to simulate or to represent the different **observed phenomena**, we need to define and describe them. To understand the difference between an observed phenomenon and the classical light, shape, and matter decomposition, we can take the example of a highlight. Its observed shape (by a human user or a sensor) is the resulting process of the interaction of these three components, and can be simulated this way. However, this does not provide any intuitive understanding of their relative influence on the final shape: an artist will directly describe the resulting shape, and not each of the three properties. We thus want to decompose the observed signal into models for each scale that can be easily understandable, representable,

and manipulable. For this purpose, we will rely on the **analysis** of the resulting interaction of light, shape, and matter as observed by a human or a physical sensor. We first consider this analysis from an **optical point of view**, trying to identify the different phenomena and their scale according to their mathematical properties (e.g., differential [70] and frequency analysis [38]). Such an approach has leaded us to exhibit the influence of surfaces flows (depth and normal gradients) into lighting pattern deformation (see Figure 4). For a **human observer**, this correspond to one recent trend in computer graphics that takes into account the human visual systems [39] both to evaluate the results and to guide the simulations.

3.4. Axis 2: From Acquisition to Display

Challenge: Convergence of optical and digital systems to blend real and virtual worlds.

Results: Instruments to acquire real world, to display virtual world, and to make both of them interact.

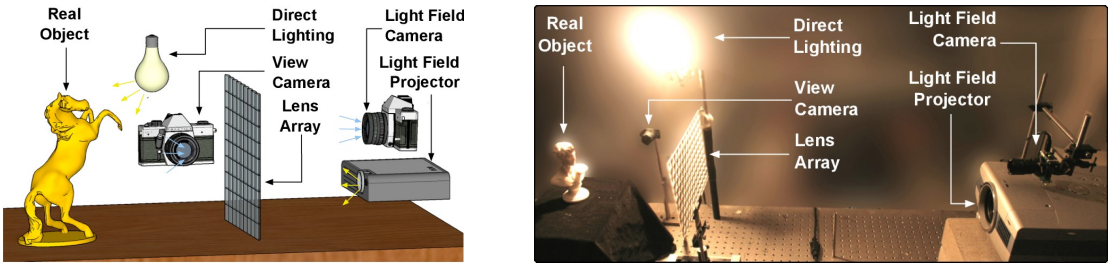


Figure 5. Light-Field transfer: global illumination between real and synthetic objects [31]

In this axis, we investigate *unified acquisition and display systems*, that is systems which combine optical instruments with digital processing. From digital to real, we investigate new display approaches [60], [47]. We consider projecting systems and surfaces [27], for personal use, virtual reality and augmented reality [22]. From the real world to the digital world, we favor direct measurements of parameters for models and representations, using (new) optical systems unless digitization is required [41], [40]. These resulting systems have to acquire the different phenomena described in Axis 1 and to display them, in an efficient manner [45], [21], [46], [49]. By efficient, we mean that we want to shorten the path between the real world and the virtual world by increasing the data bandwidth between the real (analog) and the virtual (digital) worlds, and by reducing the latency for real-time interactions (we have to prevent unnecessary conversions, and to reduce processing time). To reach this goal, the systems have to be designed as a whole, not by a simple concatenation of optical systems and digital processes, nor by considering each component independently [50].

To increase data bandwidth, one solution is to **parallelize more and more the physical systems**. One possible solution is to multiply the number of simultaneous acquisitions (e.g., simultaneous images from multiple viewpoints [49], [68]). Similarly, increasing the number of viewpoints is a way toward the creation of full 3D displays [60]. However, full acquisition or display of 3D real environments theoretically requires a continuous field of viewpoints, leading to huge data size. Despite the current belief that the increase of computational power will fill the missing gap, when it comes to visual or physical realism, if you double the processing power, people may want four times more accuracy, thus increasing data size as well. To reach the best performances, a trade-off has to be found between the amount of data required to represent accurately the reality and the amount of required processing. This trade-off may be achieved using **compressive sensing**. Compressive sensing is a new trend issued from the applied mathematics community that provides tools to accurately reconstruct a signal from a small set of measurements assuming that it is sparse in a transform domain (e.g., [67], [92]).

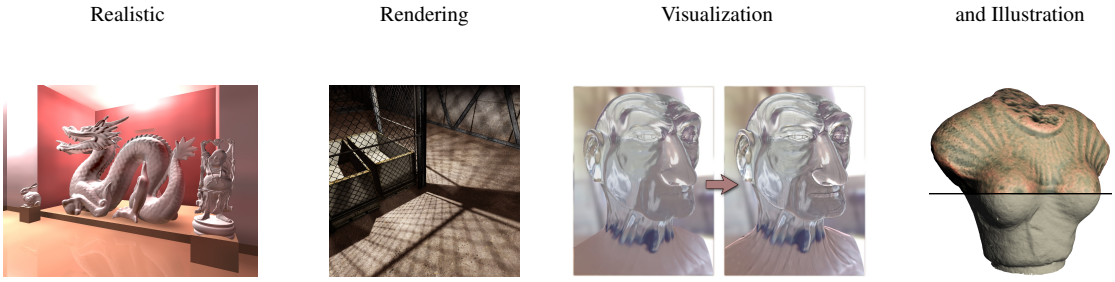
We prefer to achieve this goal by avoiding as much as possible the classical approach where acquisition is followed by a fitting step: this requires in general a large amount of measurements and the fitting itself may consume consequently too much memory and preprocessing time. By **preventing unnecessary conversion** through fitting techniques, such an approach increase the speed and reduce the data transfer for acquisition but also for display. One of the best recent examples is the work of Cossairt et al. [31]. The whole system is designed around a unique representation of the energy-field issued from (or leaving) a 3D object, either virtual or real: the Light-Field. A Light-Field encodes the light emitted in any direction from any position on an object. It is acquired thanks to a lens-array that leads to the capture of, and projection from, multiple simultaneous viewpoints. A unique representation is used for all the steps of this system. Lens-arrays, parallax barriers, and coded-aperture [57] are one of the key technologies to develop such acquisition (e.g., Light-Field camera [0] [50] and acquisition of light-sources [41]), projection systems (e.g., auto-stereoscopic displays). Such an approach is versatile and may be applied to improve classical optical instruments [55]. More generally, by designing unified optical and digital systems [64], it is possible to leverage the requirement of processing power, the memory footprint, and the cost of optical instruments.

Those are only some examples of what we investigate. We also consider the following approaches to develop new unified systems. First, similar to (and based on) the analysis goal of Axis 1, we have to take into account as much as possible the characteristics of the measurement setup. For instance, when fitting cannot be avoided, integrating them may improve both the processing efficiency and accuracy [66]. Second, we have to integrate signals from multiple sensors (such as GPS, accelerometer, ...) to prevent some computation (e.g., [58]). Finally, the experience of the group in surface modeling help the design of optical surfaces [53] for light sources or head-mounted displays.

3.5. Axis 3: Rendering, Visualization and Illustration

Challenge: How to offer the most legible signal to the final observer in real-time?

Results: High-level shading primitives, expressive rendering techniques for object depiction, real-time realistic rendering algorithms



(a) Global illumination [65] (b) Shadows [93] (c) Shape enhancement [89] (d) Shape depiction [19]

Figure 6. In the MANAO project, we are investigating rendering techniques from realistic solutions (e.g., inter-reflections (a) and shadows (b)) to more expressive ones (shape enhancement (c) with realistic style and shape depiction (d) with stylized style) for visualization.

The main goal of this axis is to offer to the final observer, in this case mostly a human user, the most legible signal in real-time. Thanks to the analysis and to the decomposition in different phenomena resulting from interactions between light, shape, and matter (Axis 1), and their perception, we can use them to convey essential information in the most pertinent way. Here, the word *pertinent* can take various forms depending on the application.

[0] Lytro, http://www.lytro.com/

In the context of scientific illustration and visualization, we are primarily interested in tools to convey shape or material characteristics of objects in animated 3D scenes. **Expressive rendering** techniques (see Figure 6 c,d) provide means for users to depict such features with their own style. To introduce our approach, we detail it from a shape-depiction point of view, domain where we have acquired a recognized expertise. Prior work in this area mostly focused on stylization primitives to achieve line-based rendering [90], [54] or stylized shading [25], [89] with various levels of abstraction. A clear representation of important 3D **object features** remains a major challenge for better shape depiction, stylization and abstraction purposes. Most existing representations provide only local properties (e.g., curvature), and thus lack characterization of broader shape features. To overcome this limitation, we are developing higher level descriptions of shape [18] with increased robustness to sparsity, noise, and outliers. This is achieved in close collaboration with Axis 1 by the use of higher-order local fitting methods, multi-scale analysis, and global regularization techniques. In order not to neglect the observer and the material characteristics of the objects, we couple this approach with an analysis of the appearance model. To our knowledge, this is an approach which has not been considered yet. This research direction is at the heart of the *MANAO* project, and has a strong connection with the analysis we plan to conduct in Axis 1. Material characteristics are always considered at the light ray level, but an understanding of **higher-level primitives** (like the shape of highlights and their motion) would help us to produce more legible renderings and permit novel stylizations; for instance, there is no method that is today able to create stylized renderings that follow the motion of highlights or shadows. We also believe such tools also play a fundamental role for geometry processing purposes (such as shape matching, reassembly, simplification), as well as for editing purposes as discussed in Axis 4.

In the context of **real-time photo-realistic rendering** ((see Figure 6 a,b), the challenge is to compute the most plausible images with minimal effort. During the last decade, a lot of work has been devoted to design approximate but real-time rendering algorithms of complex lighting phenomena such as soft-shadows [91], motion blur [38], depth of field [79], reflexions, refractions, and inter-reflexions. For most of these effects it becomes harder to discover fundamentally new and faster methods. On the other hand, we believe that significant speedup can still be achieved through more clever use of **massively parallel architectures** of the current and upcoming hardware, and/or through more clever tuning of the current algorithms. In particular, regarding the second aspect, we remark that most of the proposed algorithms depend on several parameters which can be used to **trade the speed over the quality**. Significant speed-up could thus be achieved by identifying effects that would be masked or facilitated and thus devote appropriate computational resources to the rendering [56], [37]. Indeed, the algorithm parameters controlling the quality vs speed are numerous without a direct mapping between their values and their effect. Moreover, their ideal values vary over space and time, and to be effective such an auto-tuning mechanism has to be extremely fast such that its cost is largely compensated by its gain. We believe that our various work on the analysis of the appearance such as in Axis 1 could be beneficial for such purpose too.

Realistic and real-time rendering is closely related to Axis 2: real-time rendering is a requirement to close the loop between real world and digital world. We have to thus develop algorithms and rendering primitives that allow the integration of the acquired data into real-time techniques. We have also to take care of that these real-time techniques have to work with new display systems. For instance, stereo, and more generally multi-view displays are based on the multiplication of simultaneous images. Brute force solutions consist in independent rendering pipeline for each viewpoint. A more energy-efficient solution would take advantages of the computation parts that may be factorized. Another example is the rendering techniques based on image processing, such as our work on augmented reality [29]. Independent image processing for each viewpoint may disturb the feeling of depth by introducing inconsistent information in each images. Finally, more dedicated displays [47] would require new rendering pipelines.

3.6. Axis 4: Editing and Modeling

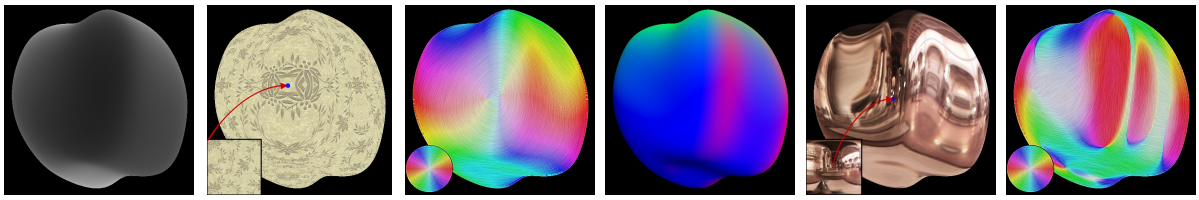
Challenge: Editing and modeling appearance using drawing- or sculpting-like tools through high level representations.

Results: High-level primitives and hybrid representations for appearance and shape.

During the last decade, the domain of computer graphics has exhibited tremendous improvements in image quality, both for 2D applications and 3D engines. This is mainly due to the availability of an ever increasing amount of shape details, and sophisticated appearance effects including complex lighting environments. Unfortunately, with such a growth in visual richness, even so-called *vectorial* representations (e.g., subdivision surfaces, Bézier curves, gradient meshes, etc.) become very dense and unmanageable for the end user who has to deal with a huge mass of control points, color labels, and other parameters. This is becoming a major challenge, with a necessity for novel representations. This Axis is thus complementary of Axis 3: the focus is the development of primitives that are easy to use for modeling and editing.

More specifically, we plan to investigate *vectorial representations* that would be amenable to the production of rich shapes with a minimal set of primitives and/or parameters. To this end we plan to build upon our insights on dynamic local reconstruction techniques and implicit surfaces [30] [24]. When working in 3D, an interesting approach to produce detailed shapes is by means of procedural geometry generation. For instance, many natural phenomena like waves or clouds may be modeled using a combination of procedural functions. Turning such functions into triangle meshes (main rendering primitives of GPUs) is a tedious process that appears not to be necessary with an adapted vectorial shape representation where one could directly turn procedural functions into implicit geometric primitives. Since we want to prevent unnecessary conversions in the whole pipeline (here, between modeling and rendering steps), we will also consider *hybrid representations* mixing meshes and implicit representations. Such research has thus to be conducted while considering the associated editing tools as well as performance issues. It is indeed important to keep *real-time performance* (cf. Axis 2) throughout the interaction loop, from user inputs to display, via editing and rendering operations. Finally, it would be interesting to add *semantic information* into 2D or 3D geometric representations. Semantic geometry appears to be particularly useful for many applications such as the design of more efficient manipulation and animation tools, for automatic simplification and abstraction, or even for automatic indexing and searching. This constitutes a complementary but longer term research direction.

In the *MANAO* project, we want to investigate representations beyond the classical light, shape, and matter decomposition. We thus want to directly control the appearance of objects both in 2D and 3D applications (e.g., [84]): this is a core topic of computer graphics. When working with 2D vector graphics, digital artists must carefully set up color gradients and textures: examples range from the creation of 2D logos to the photo-realistic imitation of object materials. Classic vector primitives quickly become impractical for creating illusions of complex materials and illuminations, and as a result an increasing amount of time and skill is required. This is only for still images. For animations, vector graphics are only used to create legible appearances composed of simple lines and color gradients. There is thus a need for more complex primitives that are able to accommodate complex reflection or texture patterns, while keeping the ease of use of vector graphics. For instance, instead of drawing color gradients directly, it is more advantageous to draw flow lines that represent local surface concavities and convexities. Going through such an intermediate structure then allows to deform simple material gradients and textures in a coherent way (see Figure 7), and animate them all at once. The manipulation of 3D object materials also raises important issues. Most existing material models are tailored to faithfully reproduce physical behaviors, not to be *easily controllable* by artists. Therefore artists learn to tweak model parameters to satisfy the needs of a particular shading appearance, which can quickly become cumbersome as the complexity of a 3D scene increases. We believe that an alternative approach is required, whereby material appearance of an object in a typical lighting environment is directly input (e.g., painted or drawn), and adapted to match a plausible material behavior. This way, artists will be able to create their own appearance (e.g., by using our shading primitives [84]), and replicate it to novel illumination environments and 3D models. For this purpose, we will rely on the decompositions and tools issued from Axis 1.



(a) (b) (c) (d) (e) (f)

Figure 7. Based on our analysis [87] (Axis 1), we have designed a system that mimics texture (left) and shading (right) effects using image processing alone. It takes depth (a) and normal (d) images as input, and uses them to deform images (b-e) in ways that closely approximate surface flows (c-f). It provides a convincing, yet artistically controllable illusion of 3D shape conveyed through texture or shading cues.

MAVERICK Project-Team

3. Research Program

3.1. Introduction

The Maverick project-team aims at producing representations and algorithms for efficient, high-quality computer generation of pictures and animations through the study of four **research problems**:

- *Computer Visualization* where we take as input a large localized dataset and represent it in a way that will let an observer understand its key properties. Visualization can be used for data analysis, for the results of a simulation, for medical imaging data...
- *Expressive Rendering*, where we create an artistic representation of a virtual world. Expressive rendering corresponds to the generation of drawings or paintings of a virtual scene, but also to some areas of computational photography, where the picture is simplified in specific areas to focus the attention.
- *Illumination Simulation*, where we model the interaction of light with the objects in the scene, resulting in a photorealistic picture of the scene. Research include improving the quality and photorealism of pictures, including more complex effects such as depth-of-field or motion-blur. We are also working on accelerating the computations, both for real-time photorealistic rendering and offline, high-quality rendering.
- *Complex Scenes*, where we generate, manage, animate and render highly complex scenes, such as natural scenes with forests, rivers and oceans, but also large datasets for visualization. We are especially interested in interactive visualization of complex scenes, with all the associated challenges in terms of processing and memory bandwidth.

The fundamental research interest of Maverick is first, *understanding* what makes a picture useful, powerful and interesting for the user, and second *designing* algorithms to create and improve these pictures.

3.2. Research approaches

We will address these research problems through three interconnected research approaches:

3.2.1. Picture Impact

Our first research axis deals with the *impact* pictures have on the viewer, and how we can improve this impact. Our research here will target:

- *evaluating user response:* we need to evaluate how the viewers respond to the pictures and animations generated by our algorithms, through user studies, either asking the viewer about what he perceives in a picture or measuring how his body reacts (eye tracking, position tracking).
- *removing artefacts and discontinuities:* temporal and spatial discontinuities perturb viewer attention, distracting the viewer from the main message. These discontinuities occur during the picture creation process; finding and removing them is a difficult process.

3.2.2. Data Representation

The data we receive as input for picture generation is often unsuitable for interactive high-quality rendering: too many details, no spatial organisation... Similarly the pictures we produce or get as input for other algorithms can contain superfluous details.

One of our goals is to develop new data representations, adapted to our requirements for rendering. This includes fast access to the relevant information, but also access to the specific hierarchical level of information needed: we want to organize the data in hierarchical levels, pre-filter it so that sampling at a given level also gives information about the underlying levels. Our research for this axis include filtering, data abstraction, simplification and stylization.

The input data can be of any kind: geometric data, such as the model of an object, scientific data before visualization, pictures and photographs. It can be time-dependent or not; time-dependent data bring an additional level of challenge on the algorithm for fast updates.

3.2.3. Prediction and simulation

Our algorithms for generating pictures require computations: sampling, integration, simulation... These computations can be optimized if we already know the characteristics of the final picture. Our recent research has shown that it is possible to predict the local characteristics of a picture by studying the phenomena involved: the local complexity, the spatial variations, their direction...

Our goal is to develop new techniques for predicting the properties of a picture, and to adapt our image-generation algorithms to these properties, for example by sampling less in areas of low variation.

Our research problems and approaches are all cross-connected. Research on the *impact* of pictures is of interest in three different research problems: *Computer Visualization*, *Expressive rendering* and *Illumination Simulation*. Similarly, our research on *Illumination simulation* will use all three research approaches: impact, representations and prediction.

3.3. Cross-cutting research issues

Beyond the connections between our problems and research approaches, we are interested in several issues, which are present throughout all our research:

sampling is an ubiquitous process occurring in all our application domains, whether photorealistic rendering (*e.g.* photon mapping), expressive rendering (*e.g.* brush strokes), texturing, fluid simulation (Lagrangian methods), etc. When sampling and reconstructing a signal for picture generation, we have to ensure both coherence and homogeneity. By *coherence*, we mean not introducing spatial or temporal discontinuities in the reconstructed signal. By *homogeneity*, we mean that samples should be placed regularly in space and time. For a time-dependent signal, these requirements are conflicting with each other, opening new areas of research.

filtering is another ubiquitous process, occuring in all our application domains, whether in realistic rendering (*e.g.* for integrating height fields, normals, material properties), expressive rendering (*e.g.* for simplifying strokes), textures (through non-linearity and discontinuities). It is especially relevant when we are replacing a signal or data with a lower resolution (for hierarchical representation); this involves filtering the data with a reconstruction kernel, representing the transition between levels.

performance and scalability are also a common requirement for all our applications. We want our algorithms to be usable, which implies that they can be used on large and complex scenes, placing a great importance on scalability. For some applications, we target interactive and real-time applications, with an update frequency between 10 Hz and 120 Hz.

coherence and continuity in space and time is also a common requirement of realistic as well as expressive models which must be ensured despite contradictory requirements. We want to avoid flickering and aliasing.

animation: our input data is likely to be time-varying (*e.g.* animated geometry, physical simulation, time-dependent dataset). A common requirement for all our algorithms and data representation is that they must be compatible with animated data (fast updates for data structures, low latency algorithms...).

3.4. Methodology

Our research is guided by several methodological principles:

Experimentation: to find solutions and phenomenological models, we use experimentation, performing statistical measurements of how a system behaves. We then extract a model from the experimental data.

Validation: for each algorithm we develop, we look for experimental validation: measuring the behavior of the algorithm, how it scales, how it improves over the state-of-the-art... We also compare our algorithms to the exact solution. Validation is harder for some of our research domains, but it remains a key principle for us.

Reducing the complexity of the problem: the equations describing certain behaviors in image synthesis can have a large degree of complexity, precluding computations, especially in real time. This is true for physical simulation of fluids, tree growth, illumination simulation... We are looking for *emerging phenomena* and *phenomenological models* to describe them (see framed box "Emerging phenomena"). Using these, we simplify the theoretical models in a controlled way, to improve user interaction and accelerate the computations.

Transferring ideas from other domains: Computer Graphics is, by nature, at the interface of many research domains: physics for the behavior of light, applied mathematics for numerical simulation, biology, algorithmics... We import tools from all these domains, and keep looking for new tools and ideas.

Develop new fondamental tools: In situations where specific tools are required for a problem, we will proceed from a theoretical framework to develop them. These tools may in return have applications in other domains, and we are ready to disseminate them.

Collaborate with industrial partners: we have a long experience of collaboration with industrial partners. These collaborations bring us new problems to solve, with short-term or medium-term transfer opportunities. When we cooperate with these partners, we have to find *what they need*, which can be very different from *what they want*, their expressed need.

MFX Project-Team

3. Research Program

3.1. Research Program

We focus on the computational aspects of shape modeling and processing for digital fabrication: dealing with shape complexity, revisiting design and customization of existing parts in view of the novel possibilities afforded by AM, and providing a stronger integration between modeling and the capabilities of the target processes.
We tackle on the following challenges:

- develop **novel shape synthesis and shape completion algorithms** that can help users model shapes with features in the scale of microns to meters, while following functional, structural, geometric and fabrication requirements;
- propose methodologies to help *expert* designers **describe shapes** and designs that can later be **customized and adapted** to different use cases;
- develop novel algorithms to **adapt and prepare complex designs** for fabrication in a given technology, including the possibility to modify aspects of the design while preserving its functionality;
- develop novel techniques to **unlock the full potential of fabrication processes**, improving their versatility in terms of feasible shapes as well as their capabilities in terms of accuracy and quality of deposition;
- develop **novel shape representations, data-structures, visualization and interaction techniques** to support the integration of our approaches into a single, unified software framework that covers the full chain from modeling to printing instructions;
- **integrate novel capabilities** enabled by advances in additive manufacturing processes and materials **in the modeling and processing chains**, in particular regarding the use of functional materials (*e.g.* piezoelectric, conductive, shrinkable).

Our approach is to cast a holistic view on the aforementioned challenges, by considering modeling and fabrication as a single, unified process. Thus, the modeling techniques we seek to develop will take into account the geometric constraints imposed by the manufacturing processes (minimal thickness, overhang angles, trapped material) as well as the desired object functionality (rigidity, porosity). To allow for the modeling of complex shapes, and to adapt the same initial design to different technologies, we propose to develop techniques that can automatically synthesize functional details within parts. At the same time, we will explore ways to increase the versatility of the manufacturing processes, through algorithms that are capable of exploiting additional degrees of freedom (*e.g.*, curved layering [21]), can introduce new capabilities (*e.g.*, material mixing [22]) and improve part accuracy (*e.g.*, adaptive slicing [20]).

Our research program is organized along three main research directions. The first one focuses on the automatic synthesis of shapes with intricate multi-scale geometries, that conform to the constraints of additive manufacturing technologies. The second direction considers geometric and algorithmic techniques for the actual fabrication of the modeled object. We aim to further improve the capabilities of the manufacturing processes with novel deposition strategies. The third direction focuses on computational design algorithms to help model parts with gradient of properties, as well as to help customizing existing designs for their reuse.

These three research directions interact strongly, and cross-pollinate: *e.g.*, novel possibilities in manufacturing unlock novel possibilities in terms of shapes that can be synthesized. Stronger synthesis methods allow for further customization.

MIMETIC Project-Team

3. Research Program

3.1. Biomechanics and Motion Control

Human motion control is a highly complex phenomenon that involves several layered systems, as shown in Figure 3 . Each layer of this controller is responsible for dealing with perceptual stimuli in order to decide the actions that should be applied to the human body and his environment. Due to the intrinsic complexity of the information (internal representation of the body and mental state, external representation of the environment) used to perform this task, it is almost impossible to model all the possible states of the system. Even for simple problems, there generally exists an infinity of solutions. For example, from the biomechanical point of view, there are much more actuators (i.e. muscles) than degrees of freedom leading to an infinity of muscle activation patterns for a unique joint rotation. From the reactive point of view there exists an infinity of paths to avoid a given obstacle in navigation tasks. At each layer, the key problem is to understand how people select one solution among these infinite state spaces. Several scientific domains have addressed this problem with specific points of view, such as physiology, biomechanics, neurosciences and psychology.

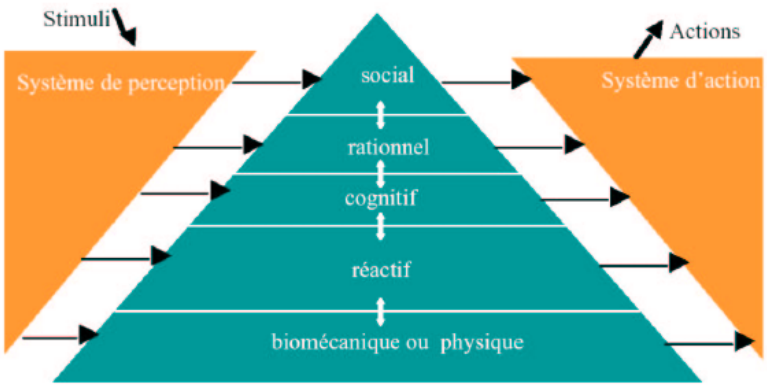


Figure 3. Layers of the motion control natural system in humans.

In biomechanics and physiology, researchers have proposed hypotheses based on accurate joint modeling (to identify the real anatomical rotational axes), energy minimization, force and torques minimization, comfort maximization (i.e. avoiding joint limits), and physiological limitations in muscle force production. All these constraints have been used in optimal controllers to simulate natural motions. The main problem is thus to define how these constraints are composed altogether such as searching the weights used to linearly combine these criteria in order to generate a natural motion. Musculoskeletal models are stereotyped examples for which there exists an infinity of muscle activation patterns, especially when dealing with antagonist muscles. An unresolved problem is to define how to use the above criteria to retrieve the actual activation patterns, while optimization approaches still leads to unrealistic ones. It is still an open problem that will require multidisciplinary skills including computer simulation, constraint solving, biomechanics, optimal control, physiology and neuroscience.

In neuroscience, researchers have proposed other theories, such as coordination patterns between joints driven by simplifications of the variables used to control the motion. The key idea is to assume that instead of controlling all the degrees of freedom, people control higher level variables which correspond to combinations of joint angles. In walking, data reduction techniques such as Principal Component Analysis have shown that lower-limb joint angles are generally projected on a unique plane whose angle in the state space is associated with energy expenditure. Although knowledge exists for specific motions, such as locomotion or grasping, this type of approach is still difficult to generalize. The key problem is that many variables are coupled and it is very difficult to objectively study the behavior of a unique variable in various motor tasks. Computer simulation is a promising method to evaluate such type of assumptions as it enables to accurately control all the variables and to check if it leads to natural movements.

Neuroscience also addresses the problem of coupling perception and action by providing control laws based on visual cues (or any other senses), such as determining how the optical flow is used to control direction in navigation tasks, while dealing with collision avoidance or interception. Coupling of the control variables is enhanced in this case as the state of the body is enriched by the large amount of external information that the subject can use. Virtual environments inhabited with autonomous characters whose behavior is driven by motion control assumptions is a promising approach to solve this problem. For example, an interesting problem in this field is navigation in an environment inhabited with other people. Typically, avoiding static obstacles together with other people displacing into the environment is a combinatory problem that strongly relies on the coupling between perception and action.

One of the main objectives of MimeTIC is to enhance knowledge on human motion control by developing innovative experiments based on computer simulation and immersive environments. To this end, designing experimental protocols is a key point and some of the researchers in MimeTIC have developed this skill in biomechanics and perception-action coupling. Associating these researchers to experts in virtual human simulation, computational geometry and constraints solving enable us to contribute to enhance fundamental knowledge in human motion control.

3.2. Experiments in Virtual Reality

Understanding interactions between humans is challenging because it addresses many complex phenomena including perception, decision-making, cognition and social behaviors. Moreover, all these phenomena are difficult to isolate in real situations, and it is therefore highly complex to understand their individual influence on these human interactions. It is then necessary to find an alternative solution that can standardize the experiments and that allows the modification of only one parameter at a time. Video was first used since the displayed experiment is perfectly repeatable and cut-offs (stop the video at a specific time before its end) allow having temporal information. Nevertheless, the absence of adapted viewpoint and stereoscopic vision does not provide depth information that are very meaningful. Moreover, during video recording session, the real human is acting in front of a camera and not of an opponent. The interaction is then not a real interaction between humans.

Virtual Reality (VR) systems allow full standardization of the experimental situations and the complete control of the virtual environment. It is then possible to modify only one parameter at a time and to observe its influence on the perception of the immersed subject. VR can then be used to understand what information is picked up to make a decision. Moreover, cut-offs can also be used to obtain temporal information about when information is picked up. When the subject can moreover react as in a real situation, his movement (captured in real time) provides information about his reactions to the modified parameter. Not only is the perception studied, but the complete perception-action loop. Perception and action are indeed coupled and influence each other as suggested by Gibson in 1979.

Finally, VR allows the validation of virtual human models. Some models are indeed based on the interaction between the virtual character and the other humans, such as a walking model. In that case, there are two ways to validate it. First, they can be compared to real data (e.g. real trajectories of pedestrians). But such data are not always available and are difficult to get. The alternative solution is then to use VR. The validation of the realism of the model is then done by immersing a real subject in a virtual environment in which a virtual

character is controlled by the model. Its evaluation is then deduced from how the immersed subject reacts when interacting with the model and how realistic he feels the virtual character is.

3.3. Computer animation

Computer animation is the branch of computer science devoted to models for the representation and simulation of the dynamic evolution of virtual environments. A first focus is the animation of virtual characters (behavior and motion). Through a deeper understanding of interactions using VR and through better perceptive, biomechanical and motion control models to simulate the evolution of dynamic systems, the Mimetic team has the ability to build more realistic, efficient and believable animations. Perceptual study also enables us to focus computation time on relevant information (i.e. leading to ensure natural motion from the perceptual points of view) and save time for unperceived details. The underlying challenges are (i) the computational efficiency of the system which needs to run in real-time in many situations, (ii) the capacity of the system to generalise/adapt to new situations for which data was not available or for which models were not defined for, and (iii) the variability of the models, i.e. their ability to handle many body morphologies and generate variations in motions that would be specific to each virtual character.

In many cases, however, these challenges cannot be addressed in isolation. Typically character behaviors also depend on the nature and the topology of the environment they are surrounded by. In essence, a character animation system should also rely on smarter representations of the environments, in order to better perceive the environment, and take contextualised decisions. Hence the animation of virtual characters in our context often requires to be coupled with models to represent the environment, reason, and plan both at a geometric level (can the character reach this location), and at a semantic level (should it use the sidewalk, the stairs, or the road). This represents the second focus. Underlying challenges are the ability to offer a compact, yet precise representation on which efficient path and motion planning can be performed, and on which high-level reasonning can be achieved.

Finally, a third scientific focus tied to the computer animation axis is digital storytelling. Evolved representations of motions and environments enable realistic animations. It is yet equally important to question how these event should be portrayed, when and under which angle. In essence, this means integrating *discourse models* into *story models*, the story representing the sequence of events which occur in a virtual environment, and the discourse representing how this story should be displayed (ie which events to show in which order and with which viewpoint). Underlying challenges are pertained to (i) narrative discourse representations, (ii) projections of the discourse into the geometry, planning camera trajectories and planning cuts between the viewpoints and (iii) means to interactively control the unfolding of the discourse.

By therefore establishing the foundations to build bridges between the high-level narrative structures, the semantic/geometric planning of motions and events, and low-level character animations, the Mimetic team adopts a principled and all-inclusive approach to the animation of virtual characters.

POTIOC Project-Team

3. Research Program

3.1. Research Program

To achieve our overall objective, we follow two main research axes, plus one transverse axis, as illustrated in Figure 2 .

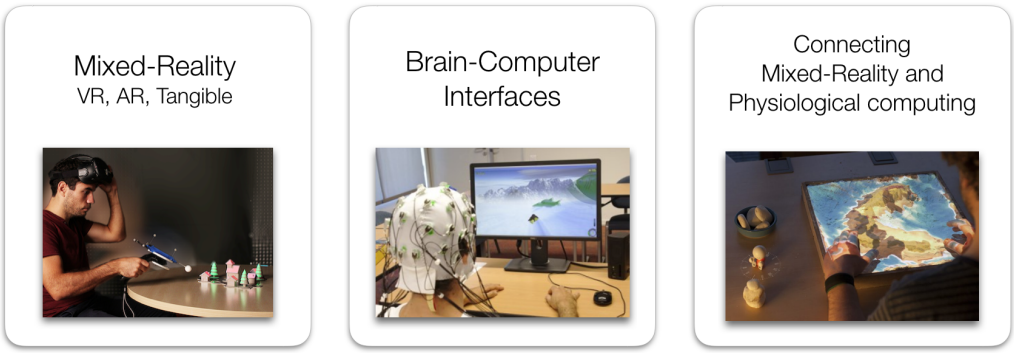


Figure 2. Main research axes of Potioc.

In the first axis dedicated to **Interaction in Mixed-Reality spaces**, we explore interaction paradigms that encompass virtual and/or physical objects. We are notably interested in hybrid environments that co-locate virtual and physical spaces, and we also explore approaches that allow one to move from one space to the other.

The second axis is dedicated to **Brain-Computer Interfaces (BCI)**, i.e., systems enabling user to interact by means of brain activity only. We target BCI systems that are reliable and accessible to a large number of people. To do so, we work on brain signal processing algorithms as well as on understanding and improving the way we train our users to control these BCIs.

Finally, in the **transverse** axis, we explore new approaches that involve both mixed-reality and neuro-physiological signals. In particular, tangible and augmented objects allow us to explore interactive physical visualizations of human inner states. Physiological signals also enable us to better assess user interaction, and consequently, to refine the proposed interaction techniques and metaphors.

From a methodological point of view, for these three axes, we work at three different interconnected levels. The first level is centered on the human sensori-motor and cognitive abilities, as well as user strategies and preferences, for completing interaction tasks. We target, in a fundamental way, a better understanding of humans interacting with interactive systems. The second level is about the creation of interactive systems. This notably includes development of hardware and software components that will allow us to explore new input and output modalities, and to propose adapted interaction techniques. Finally, in a last higher level, we are interested in specific application domains. We want to contribute to the emergence of new applications and usages, with a societal impact.

TITANE Project-Team

3. Research Program

3.1. Context

Geometric modeling and processing revolve around three main end goals: a computerized shape representation that can be visualized (creating a realistic or artistic depiction), simulated (anticipating the real) or realized (manufacturing a conceptual or engineering design). Aside from the mere editing of geometry, central research themes in geometric modeling involve conversions between physical (real), discrete (digital), and mathematical (abstract) representations. Going from physical to digital is referred to as shape acquisition and reconstruction; going from mathematical to discrete is referred to as shape approximation and mesh generation; going from discrete to physical is referred to as shape rationalization.

Geometric modeling has become an indispensable component for computational and reverse engineering. Simulations are now routinely performed on complex shapes issued not only from computer-aided design but also from an increasing amount of available measurements. The scale of acquired data is quickly growing: we no longer deal exclusively with individual shapes, but with entire *scenes*, possibly at the scale of entire cities, with many objects defined as structured shapes. We are witnessing a rapid evolution of the acquisition paradigms with an increasing variety of sensors and the development of community data, as well as disseminated data.

In recent years, the evolution of acquisition technologies and methods has translated in an increasing overlap of algorithms and data in the computer vision, image processing, and computer graphics communities. Beyond the rapid increase of resolution through technological advances of sensors and methods for mosaicing images, the line between laser scan data and photos is getting thinner. Combining, e.g., laser scanners with panoramic cameras leads to massive 3D point sets with color attributes. In addition, it is now possible to generate dense point sets not just from laser scanners but also from photogrammetry techniques when using a well-designed acquisition protocol. Depth cameras are getting increasingly common, and beyond retrieving depth information we can enrich the main acquisition systems with additional hardware to measure geometric information about the sensor and improve data registration: e.g., accelerometers or GPS for geographic location, and compasses or gyrometers for orientation. Finally, complex scenes can be observed at different scales ranging from satellite to pedestrian through aerial levels.

These evolutions allow practitioners to measure urban scenes at resolutions that were until now possible only at the scale of individual shapes. The related scientific challenge is however more than just dealing with massive data sets coming from increase of resolution, as complex scenes are composed of multiple objects with structural relationships. The latter relate i) to the way the individual shapes are grouped to form objects, object classes or hierarchies, ii) to geometry when dealing with similarity, regularity, parallelism or symmetry, and iii) to domain-specific semantic considerations. Beyond reconstruction and approximation, consolidation and synthesis of complex scenes require rich structural relationships.

The problems arising from these evolutions suggest that the strengths of geometry and images may be combined in the form of new methodological solutions such as photo-consistent reconstruction. In addition, the process of measuring the geometry of sensors (through gyrometers and accelerometers) often requires both geometry process and image analysis for improved accuracy and robustness. Modeling urban scenes from measurements illustrates this growing synergy, and it has become a central concern for a variety of applications ranging from urban planning to simulation through rendering and special effects.

3.2. Analysis

Complex scenes are usually composed of a large number of objects which may significantly differ in terms of complexity, diversity, and density. These objects must be identified and their structural relationships must be recovered in order to model the scenes with improved robustness, low complexity, variable levels of details and ultimately, semantization (automated process of increasing degree of semantic content).

Object classification is an ill-posed task in which the objects composing a scene are detected and recognized with respect to predefined classes, the objective going beyond scene segmentation. The high variability in each class may explain the success of the stochastic approach which is able to model widely variable classes. As it requires a priori knowledge this process is often domain-specific such as for urban scenes where we wish to distinguish between instances as ground, vegetation and buildings. Additional challenges arise when each class must be refined, such as roof super-structures for urban reconstruction.

Structure extraction consists in recovering structural relationships between objects or parts of object. The structure may be related to adjacencies between objects, hierarchical decomposition, singularities or canonical geometric relationships. It is crucial for effective geometric modeling through levels of details or hierarchical multiresolution modeling. Ideally we wish to learn the structural rules that govern the physical scene manufacturing. Understanding the main canonical geometric relationships between object parts involves detecting regular structures and equivalences under certain transformations such as parallelism, orthogonality and symmetry. Identifying structural and geometric repetitions or symmetries is relevant for dealing with missing data during data consolidation.

Data consolidation is a problem of growing interest for practitioners, with the increase of heterogeneous and defect-laden data. To be exploitable, such defect-laden data must be consolidated by improving the data sampling quality and by reinforcing the geometrical and structural relations sub-tending the observed scenes. Enforcing canonical geometric relationships such as local coplanarity or orthogonality is relevant for registration of heterogeneous or redundant data, as well as for improving the robustness of the reconstruction process.

3.3. Approximation

Our objective is to explore the approximation of complex shapes and scenes with surface and volume meshes, as well as on surface and domain tiling. A general way to state the shape approximation problem is to say that we search for the shape discretization (possibly with several levels of detail) that realizes the best complexity / distortion trade-off. Such a problem statement requires defining a discretization model, an error metric to measure distortion as well as a way to measure complexity. The latter is most commonly expressed in number of polygon primitives, but other measures closer to information theory lead to measurements such as number of bits or minimum description length.

For surface meshes we intend to conceive methods which provide control and guarantees both over the global approximation error and over the validity of the embedding. In addition, we seek for resilience to heterogeneous data, and robustness to noise and outliers. This would allow repairing and simplifying triangle soups with cracks, self-intersections and gaps. Another exploratory objective is to deal generically with different error metrics such as the symmetric Hausdorff distance, or a Sobolev norm which mixes errors in geometry and normals.

For surface and domain tiling the term meshing is substituted for tiling to stress the fact that tiles may be not just simple elements, but can model complex smooth shapes such as bilinear quadrangles. Quadrangle surface tiling is central for the so-called *resurfacing* problem in reverse engineering: the goal is to tile an input raw surface geometry such that the union of the tiles approximates the input well and such that each tile matches certain properties related to its shape or its size. In addition, we may require parameterization domains with a simple structure. Our goal is to devise surface tiling algorithms that are both reliable and resilient to defect-laden inputs, effective from the shape approximation point of view, and with flexible control upon the structure of the tiling.

3.4. Reconstruction

Assuming a geometric dataset made out of points or slices, the process of shape reconstruction amounts to recovering a surface or a solid that matches these samples. This problem is inherently ill-posed as infinitely-many shapes may fit the data. One must thus regularize the problem and add priors such as simplicity or smoothness of the inferred shape.

The concept of geometric simplicity has led to a number of interpolating techniques commonly based upon the Delaunay triangulation. The concept of smoothness has led to a number of approximating techniques that commonly compute an implicit function such that one of its isosurfaces approximates the inferred surface. Reconstruction algorithms can also use an explicit set of prior shapes for inference by assuming that the observed data can be described by these predefined prior shapes. One key lesson learned in the shape problem is that there is probably not a single solution which can solve all cases, each of them coming with its own distinctive features. In addition, some data sets such as point sets acquired on urban scenes are very domain-specific and require a dedicated line of research.

In recent years the *smooth, closed case* (i.e., shapes without sharp features nor boundaries) has received considerable attention. However, the state-of-the-art methods have several shortcomings: in addition to being in general not robust to outliers and not sufficiently robust to noise, they often require additional attributes as input, such as lines of sight or oriented normals. We wish to devise shape reconstruction methods which are both geometrically and topologically accurate without requiring additional attributes, while exhibiting resilience to defect-laden inputs. Resilience formally translates into stability with respect to noise and outliers. Correctness of the reconstruction translates into convergence in geometry and (stable parts of) topology of the reconstruction with respect to the inferred shape known through measurements.

Moving from the smooth, closed case to the *piecewise smooth case* (possibly with boundaries) is considerably harder as the ill-posedness of the problem applies to each sub-feature of the inferred shape. Further, very few approaches tackle the combined issue of robustness (to sampling defects, noise and outliers) and feature reconstruction.

ALMANACH Project-Team

3. Research Program

3.1. Research strands

As described above, ALMAnaCH's scientific programme is organised around three research axes. The first two aim to tackle the challenge of language variation in two complementary directions. They are supported by a third, transverse research axis on language resources. Our four-year objectives are described in much greater detail in the project-team proposal, whose very recent final validation in June 2019 resulted in the upgrade of ALMAnaCH to the "project-team" status in July 2019. They can be summarised as follows:

3.1.1. Research axis 1

Our first objective is to **stay at a state-of-the-art level in key NLP tasks** such as shallow processing, part-of-speech tagging and (syntactic) parsing, which are core expertise domains of ALMAnaCH members. This will also require us to improve the **generation of semantic representations (semantic parsing)**, and to begin to explore tasks such as machine translation, which now relies on neural architectures also used for some of the above-mentioned tasks. Given the generalisation of neural models in NLP, we will also be involved in better understanding how such models work and what they learn, something that is directly related to the investigation of language variation (Research axis 2). We will also work on the **integration of both linguistic and non-linguistic contextual information** to improve automatic linguistic analysis. This is an emerging and promising line of research in NLP. We will have to identify, model and take advantage of each type of contextual information available. Addressing these issues will enable the development of new lines of research related to conversational content. Applications include improved information and knowledge extraction algorithms. We will especially focus on challenging datasets such as domain-specific texts (e.g. financial, legal) as well as historical documents, in the larger context of the development of digital humanities. We currently also explore the even more challenging new direction of a cognitively inspired NLP, in order to tackle the possibility to enrich the architecture of state-of-the-art algorithms, such as RNNGs, based on human neuroimaging-driven data.

3.1.2. Research axis 2

Language variation must be better understood and modelled in all its forms. In this regard, we will put a strong emphasis on **four types** of language variation and their mutual interaction: **sociolinguistic variation** in synchrony (including non-canonical spelling and syntax in user-generated content), **complexity-based variation** in relation to language-related disabilities, and **diachronic variation** (computational exploration of language change and language history, with a focus on Old to all forms of Modern French, as well as Indo-European languages in general). In addition, the noise introduced by Optical Character Recognition and Handwritten Text Recognition systems, especially in the context of historical documents, bears some similarities to that of non-canonical input in user-generated content (e.g. erroneous characters). This noise constitutes a more transverse kind of variation stemming from the way language is graphically encoded, which we call **language-encoding variation**. Other types of language variation will also become important research topics for ALMAnaCH in the future. This includes dialectal variation (e.g. work on Arabic varieties, something on which we have already started working, producing the first annotated data set on Maghrebi Arabizi, the Arabic variants used on social media by people from North-African countries, written using a non-fixed Latin-script transcription) as well as the study and exploitation of paraphrases in a broader context than the above-mentioned complexity-based variation.

Both research axes above rely on the availability of language resources (corpora, lexicons), which is the focus of our third, transverse research axis.

3.1.3. Research axis 3

Language resource development (raw and annotated corpora, lexical resources) is not just a necessary preliminary step to create both evaluation datasets for NLP systems and training datasets for NLP systems based on machine learning. When dealing with datasets of interest to researchers from the humanities (e.g. large archives), it is also a goal *per se* and a preliminary step before making such datasets available and exploitable online. It involves a number of scientific challenges, among which (i) tackling issues related to the digitalisation of non-electronic datasets, (ii) tackling issues related to the fact that many DH-related datasets are domain-specific and/or not written in contemporary languages; (iii) the development of semi-automatic and automatic algorithms to speed up the work (e.g. automatic extraction of lexical information, low-resource learning for the development of pre-annotation algorithms, transfer methods to leverage existing tools and/or resources for other languages, etc.) and (iv) the development of formal models to represent linguistic information in the best possible way, thus requiring expertise at least in NLP and in typological and formal linguistics. Such endeavours are domains of expertise of the ALMAnaCH team, and a large part of our research activities will be dedicated to language resource development. In this regard, we aim to retain our leading role in the representation and management of lexical resource and treebank development and also to develop a complete processing line for the transcription, analysis and processing of complex documents of interest to the humanities, in particular archival documents. This research axis 3 will benefit the whole team and beyond, and will benefit from and feed the work of the other research axes.

3.2. Automatic Context-augmented Linguistic Analysis

This first research strand is centred around NLP technologies and some of their applications in Artificial Intelligence (AI). Core NLP tasks such as part-of-speech tagging, syntactic and semantic parsing is improved by integrating new approaches, such as (deep) neural networks, whenever relevant, while preserving and taking advantage of our expertise on symbolic and statistical system: hybridisation not only couples symbolic and statistical approaches, but neural approaches as well. AI applications are twofold, notwithstanding the impact of language variation (see the next strand): (i) information and knowledge extraction, whatever the type of input text (from financial documents to ancient, historical texts and from Twitter data to Wikipedia) and (ii) chatbots and natural language generation. In many cases, our work on these AI applications is carried out in collaboration with industrial partners. The specificities and issues caused by language variation (a text in Old French, a contemporary financial document and tweets with a non-canonical spelling cannot be processed in the same way) are addressed in the next research strand.

3.2.1. Processing of natural language at all levels: morphology, syntax, semantics

Our expertise in NLP is the outcome of more than 10 years in developing new models of analysis and accurate techniques for the full processing of any kind of language input since the early days of the Atoll project-team and the rise of linguistically informed data-driven models as put forward within the Alpage project-team.

Traditionally, a full natural language process (NLP) chain is organised as a pipeline where each stage of analysis represents a traditional linguistic field (in a *structuralism* view) from morphological analysis to purely semantic representations. The problem is that this architecture is vulnerable to error propagation and very domain sensitive: each of these stage must be compatible at the lexical and structure levels they provide. We arguably built the best performing NLP chain for French [74], [112] and one of the best for robust multilingual parsing as shown by our results in various shared tasks over the years [108], [105], [111], [82]. So we pursue our efforts on each of our components we developed: tokenisers (e.g. SxPipe), part-of-speech taggers (e.g. MElt), constituency parsers and dependency parsers (e.g. FRMG, DyALog-SR) as well as our recent neural semantic graph parsers [105].

In particular, we continue to explore the hybridisation of symbolic and statistical approaches, and extend it to neural approaches, as initiated in the context of our participation to the CoNLL 2017 multilingual parsing shared task [0] and to Extrinsic Parsing Evaluation Shared Task [0].

[0] We ranked 3 for UPOS tagging and 6 for dependency parsing out of 33 participants.

Fundamentally, we want to build tools that are less sensitive to variation, more easily configurable, and self-adapting. Our short-term goal is to explore techniques such as multi-task learning (cf. already [110]) to propose a joint model of tokenisation, normalisation, morphological analysis and syntactic analysis. We also explore adversarial learning, considering the drastic variation we face in parsing user-generated content and processing historical texts, both seen as noisy input that needs to be handled at training and decoding time.

3.2.2. Integrating context in NLP systems

While those points are fundamental, therefore necessary, if we want to build the next generation of NLP tools, we need to *push the envelop* even further by tackling the biggest current challenge in NLP: handling the context within which a speech act is taking place.

There is indeed a strong tendency in NLP to assume that each sentence is independent from its siblings sentences as well as its context of enunciation, with the obvious objective to simplify models and reduce the complexity of predictions. While this practice is already questionable when processing full-length edited documents, it becomes clearly problematic when dealing with short sentences that are noisy, full of ellipses and external references, as commonly found in User-Generated Content (UGC).

A more expressive and context-aware structural representation of a linguistic production is required to accurately model UGC. Let us consider for instance the case for Syntax-based Machine Translation of social media content, as is carried out by the ALMAnaCH-led ANR project Parsiti (PI: DS). A Facebook post may be part of a discussion thread, which may include links to external content. Such information is required for a complete representation of the post's context, and in turn its accurate machine translation. Even for the presumably simpler task of POS tagging of dialogue sequences, the addition of context-based features (namely information about the speaker and dialogue moves) was beneficial [85]. In the case of UGC, working across sentence boundaries was explored for instance, with limited success, by [73] for document-wise parsing and by [96] for POS tagging.

Taking the context into account requires new inference methods able to share information between sentences as well as new learning methods capable of finding out which information is to be made available, and where. Integrating contextual information at all steps of an NLP pipeline is among the main research questions addressed in this research strand. In the short term, we focus on morphological and syntactic disambiguation within close-world scenarios, as found in video games and domain-specific UGC. In the long term, we investigate the integration of linguistically motivated semantic information into joint learning models.

From a more general perspective, contexts may take many forms and require imagination to discern them, get useful data sets, and find ways to exploit them. A context may be a question associated with an answer, a rating associated with a comment (as provided by many web services), a thread of discussions (e-mails, social media, digital assistants, chatbots—on which see below–), but also meta data about some situation (such as discussions between gamers in relation with the state of the game) or multiple points of views (pictures and captions, movies and subtitles). Even if the relationship between a language production and its context is imprecise and indirect, it is still a valuable source of information, notwithstanding the need for less supervised machine learning techniques (cf. the use of LSTM neural networks by Google to automatically suggest replies to emails).

3.2.3. Information and knowledge extraction

The use of local contexts as discussed above is a new and promising approach. However, a more traditional notion of global context or world knowledge remains an open question and still raises difficult issues. Indeed, many aspects of language such as ambiguities and ellipsis can only be handled using world knowledge. Linked Open Data (LODs) such as DBpedia, WordNet, BabelNet, or Framebase provide such knowledge and we plan to exploit them.

[0]Semantic graph parsing, evaluated on biomedical data, speech and opinion. We ranked 1 in a joint effort with the Stanford NLP team

However, each specialised domain (economy, law, medicine...) exhibits its own set of concepts with associated terms. This is also true of communities (e.g. on social media), and it is even possible to find communities discussing the same topics (e.g. immigration) with very distinct vocabularies. Global LODs weakly related to language may be too general and not sufficient for a specific language variant. Following and extending previous work in ALPAGE, we put an emphasis on information acquisition from corpora, including error mining techniques in parsed corpora (to detect specific usages of a word that are missing in existing resources), terminology extraction, and word clustering.

Word clustering is of specific importance. It relies on the distributional hypothesis initially formulated by Harris, which states that words occurring in similar contexts tend to be semantically close. The latest developments of these ideas (with word2vec or GloVe) have led to the embedding of words (through vectors) in low-dimensional semantic spaces. In particular, words that are typical of several communities (see above) can be embedded in a same semantic space in order to establish mappings between them. It is also possible in such spaces to study static configurations and vector shifts with respect to variables such as time, using topological theories (such as pretopology), for instance to explore shifts in meaning over time (cf. the ANR project Profiterole concerning ancient French texts) or between communities (cf. the ANR project SoSweet). It is also worth mentioning on-going work (in computational semantics) whose goal is to combine word embeddings to embed expressions, sentences, paragraphs or even documents into semantic spaces, e.g. to explore the similarity of documents at various time periods.

Besides general knowledge about a domain, it is important to detect and keep trace of more specific pieces of information when processing a document and maintaining a context, especially about (recurring) Named Entities (persons, organisations, locations...) —something that is the focus of future work in collaboration with Patrice Lopez on named entity detection in scientific texts. Through the co-supervision of a PhD funded by the LabEx EFL (see below), we are also involved in pronominal coreference resolution (finding the referent of pronouns). Finally, we plan to continue working on deeper syntactic representations (as initiated with the Deep Sequoia Treebank), thus paving the way towards deeper semantic representations. Such information is instrumental when looking for more precise and complete information about who does what, to whom, when and where in a document. These lines of research are motivated by the need to extract useful contextual information, but it is also worth noting their strong potential in industrial applications.

3.3. Computational Modelling of Linguistic Variation

NLP and DH tools and resources are very often developed for contemporary, edited, non-specialised texts, often based on journalistic corpora. However, such corpora are not representative of the variety of existing textual data. As a result, the performance of most NLP systems decreases, sometimes dramatically, when faced with non-contemporary, non-edited or specialised texts. Despite the existence of domain-adaptation techniques and of robust tools, for instance for social media text processing, dealing with linguistic variation is still a crucial challenge for NLP and DH.

Linguistic variation is not a monolithic phenomenon. Firstly, it can result from different types of processes, such as variation over time (diachronic variation) and variation correlated with sociological variables (sociolinguistic variation, especially on social networks). Secondly, it can affect all components of language, from spelling (languages without a normative spelling, spelling errors of all kinds and origins) to morphology/syntax (especially in diachrony, in texts from specialised domains, in social media texts) and semantics/pragmatics (again in diachrony, for instance). Finally, it can constitute a property of the data to be analysed or a feature of the data to be generated (for instance when trying to simplify texts for increasing their accessibility for disabled and/or non-native readers).

Nevertheless, despite this variability in variation, the underlying mechanisms are partly comparable. This motivates our general vision that many generic techniques could be developed and adapted to handle different types of variation. In this regard, three aspects must be kept in mind: spelling variation (human errors, OCR/HTR errors, lack of spelling conventions for some languages...), lack or scarcity of parallel data aligning “variation-affected” texts and their “standard/edited” counterpart, and the sequential nature of the problem at hand. We will therefore explore, for instance, how unsupervised or weakly-supervised techniques

could be developed and feed dedicated sequence-to-sequence models. Such architectures could help develop "normalisation" tools adapted, for example, to social media texts, texts written in ancient/dialectal varieties of well-resourced languages (e.g. Old French texts), and OCR/HTR system outputs.

Nevertheless, the different types of language variation will require specific models, resources and tools. All these directions of research constitute the core of our second research strand described in this section.

3.3.1. Theoretical and empirical synchronic linguistics

Permanent members involved: all

We aim to explore computational models to deal with language variation. It is important to get more insights about language in general and about the way humans apprehend it. We will do so in at least two directions, associating computational linguistics with formal and descriptive linguistics on the one hand (especially at the morphological level) and with cognitive linguistics on the other hand (especially at the syntactic level).

Recent advances in morphology rely on quantitative and computational approaches and, sometimes, on collaboration with descriptive linguists—see for instance the special issue of the *Morphology* journal on "computational methods for descriptive and theoretical morphology", edited and introduced by [70]. In this regard, ALMAnaCH members have taken part in the design of quantitative approaches to defining and measuring morphological complexity and to assess the internal structure of morphological systems (inflection classes, predictability of inflected forms...). Such studies provide valuable insights on these prominent questions in theoretical morphology. They also improve the linguistic relevance and the development speed of NLP-oriented lexicons, as also demonstrated by ALMAnaCH members. We shall therefore pursue these investigations, and orientate them towards their use in diachronic models (see section 3.3.3).

Regarding cognitive linguistics, we have the perfect opportunity with the starting ANR-NSF project "Neuro-Computational Models of Natural Language" (NCM-NL) to go in this direction, by examining potential correlations between medical imagery applied on patients listening to a reading of "Le Petit Prince" and computation models applied on the novel. A secondary prospective benefit from the project will be information about processing evolution (by the patients) along the novel, possibly due to the use of contextual information by humans.

3.3.2. Sociolinguistic variation

Because language is central in our social interactions, it is legitimate to ask how the rise of digital content and its tight integration in our daily life has become a factor acting on language. This is even more actual as the recent rise of novel digital services opens new areas of expression, which support new linguistic behaviours. In particular, social media such as Twitter provide channels of communication through which speakers/writers use their language in ways that differ from standard written and oral forms. The result is the emergence of new language varieties.

A very similar situation exists with regard to historical texts, especially documentary texts or graffiti but even literary texts, that do not follow standardised orthography, morphology or syntax.

However, NLP tools are designed for standard forms of language and exhibit a drastic loss of accuracy when applied to social media varieties or non-standardised historical sources. To define appropriate tools, descriptions of these varieties are needed. However, to validate such descriptions, tools are also needed. We address this chicken-and-egg problem in an interdisciplinary fashion, by working both on linguistic descriptions and on the development of NLP tools. Recently, socio-demographic variables have been shown to bear a strong impact on NLP processing tools (see for instance [80] and references therein). This is why, in a first step, jointly with researchers involved in the ANR project SoSweet (ENS Lyon and Inria project-team Dante), we will study how these variables can be factored out by our models and, in a second step, how they can be accurately predicted from sources lacking these kinds of featured descriptions.

3.3.3. Diachronic variation

Language change is a type of variation pertaining to the diachronic axis. Yet any language change, whatever its nature (phonetic, syntactic...), results from a particular case of synchronic variation (competing phonetic realisations, competing syntactic constructions...). The articulation of diachronic and synchronic variation is influenced to a large extent by both language-internal factors (i.e. generalisation of context-specific facts) and/or external factors (determined by social class, register, domain, and other types of variation).

Very few computational models of language change have been developed. Simple deterministic finite-state-based phonetic evolution models have been used in different contexts. The PIElexicon project [90] uses such models to automatically generate forms attested in (classical) Indo-European languages but is based on an idiosyncratic and unacceptable reconstruction of the Proto-Indo-European language. Probabilistic finite-state models have also been used for automatic cognate detection and proto-form reconstruction, for example by [71] and [81]. Such models rely on a good understanding of the phonetic evolution of the languages at hand.

In ALMAnaCH, our goal is to work on modelling phonetic, morphological and lexical diachronic evolution, with an emphasis on computational etymological research and on the computational modelling of the evolution of morphological systems (morphological grammar and morphological lexicon). These efforts will be in direct interaction with sub-strand 3b (development of lexical resources). We want to go beyond the above-mentioned purely phonetic models of language and lexicon evolution, as they fail to take into account a number of crucial dimensions, among which: (1) spelling, spelling variation and the relationship between spelling and phonetics; (2) synchronic variation (geographical, genre-related, etc.); (3) morphology, especially through intra-paradigmatic and inter-paradigmatic analogical levelling phenomena, (4) lexical creation, including via affixal derivation, back-formation processes and borrowings.

We apply our models to two main tasks. The first task, as developed for example in the context of the ANR project Profiterole, consists in predicting non-attested or non-documented words at a certain date based on attestations of older or newer stages of the same word (e.g., predicting a non-documented Middle French word based on its Vulgar Latin and Old French predecessors and its Modern French successor). Morphological models and lexical diachronic evolution models will provide independent ways to perform the same predictions, thus reinforcing our hypotheses or pointing to new challenges.

The second application task is computational etymology and proto-language reconstruction. Our lexical diachronic evolution models will be paired with semantic resources (wordnets, word embeddings, and other corpus-based statistical information). This will allow us to formally validate or suggest etymological or cognate relations between lexical entries from different languages of a same language family, provided they are all inherited. Such an approach could also be adapted to include the automatic detection of borrowings from one language to another (e.g. for studying the non-inherited layers in the Ancient Greek lexicon). In the longer term, we will investigate the feasibility of the automatic (unsupervised) acquisition of phonetic change models, especially when provided with lexical data for numerous languages from the same language family.

These lines of research will rely on etymological data sets and standards for representing etymological information (see Section 3.4.2).

Diachronic evolution also applies to syntax, and in the context of the ANR project Profiterole, we are beginning to explore more or less automatic ways of detecting these evolutions and suggest modifications, relying on fine-grained syntactic descriptions (as provided by meta-grammars), unsupervised sentence clustering (generalising previous works on error mining, cf. [6]), and constraint relaxation (in meta-grammar classes). The underlying idea is that a new syntactic construction evolves from a more ancient one by small, iterative modifications, for instance by changing word order, adding or deleting functional words, etc.

3.3.4. Accessibility-related variation

Language variation does not always pertain to the textual input of NLP tools. It can also be characterised by their intended output. This is the perspective from which we investigate the issue of text simplification (for a recent survey, see for instance [109]). Text simplification is an important task for improving the accessibility to information, for instance for people suffering from disabilities and for non-native speakers learning a given

language [91]. To this end, guidelines have been developed to help writing documents that are easier to read and understand, such as the FALC ("Facile À Lire et à Comprendre") guidelines for French. [0]

Fully automated text simplification is not suitable for producing high-quality simplified texts. Besides, the involvement of disabled people in the production of simplified texts plays an important social role. Therefore, following previous works [79], [103], our goal will be to develop tools for the computer-aided simplification of textual documents, especially administrative documents. Many of the FALC guidelines can only be linguistically expressed using complex, syntactic constraints, and the amount of available "parallel" data (aligned raw and simplified documents) is limited. We will therefore investigate hybrid techniques involving rule-based, statistical and neural approaches based on parsing results (for an example of previous parsing-based work, see [68]). Lexical simplification, another aspect of text simplification [86], [92], will also be pursued. In this regard, we have already started a collaboration with Facebook's AI Research in Paris, the UNAPEI (the largest French federation of associations defending and supporting people with intellectual disabilities and their families), and the French Secretariat of State in charge of Disabled Persons.

Accessibility can also be related to the various presentation forms of a document. This is the context in which we have initiated the OPALINE project, funded by the *Programme d'Investissement d'Avenir - Fonds pour la Société Numérique*. The objective is for us to further develop the GROBID text-extraction suite [0] in order to be able to re-publish existing books or dictionaries, available in PDF, in a format that is accessible by visually impaired persons.

3.4. Modelling and Development of Language Resources

Language resources (raw and annotated corpora, lexical resources, etc.) are required in order to apply any machine learning technique (statistical, neural, hybrid) to an NLP problem, as well as to evaluate the output of an NLP system.

In data-driven, machine-learning-based approaches, language resources are the place where linguistic information is stored, be it implicitly (as in raw corpora) or explicitly (as in annotated corpora and in most lexical resources). Whenever linguistic information is provided explicitly, it complies to guidelines that formally define which linguistic information should be encoded, and how. Designing linguistically meaningful and computationally exploitable ways to encode linguistic information within language resources constitutes the first main scientific challenge in language resource development. It requires a strong expertise on both the linguistic issues underlying the type of resource under development (e.g. on syntax when developing a treebank) and the NLP algorithms that will make use of such information.

The other main challenge regarding language resource development is a consequence of the fact that it is a costly, often tedious task. ALMAnaCH members have a long track record of language resource development, including by hiring, training and supervising dedicated annotators. But a manual annotation can be speeded up by automatic techniques. ALMAnaCH members have also work on such techniques, and published work on approaches such as automatic lexical information extraction, annotation transfer from a language to closely related languages, and more generally on the use of pre-annotation tools for treebank development and on the impact of such tools on annotation speed and quality. These techniques are often also relevant for Research strand 1. For example, adapting parsers from one language to the other or developing parsers that work on more than one language (e.g. a non-lexicalised parser trained on the concatenation of treebanks from different languages in the same language family) can both improve parsing results on low-resource languages and speed up treebank development for such languages.

3.4.1. Construction, management and automatic annotation of Text Corpora

Corpus creation and management (including automatic annotation) is often a time-consuming and technically challenging task. In many cases, it also raises scientific issues related for instance with linguistic questions

[0] Please click here for an archived version of these guidelines (at the time this footnote is begin written, the original link does not seem to work any more).

[0] https://github.com/kermitt2/grobid

(what is the elementary unit in a text?) as well as computer-science challenges (for instance when OCR or HTR are involved). It is therefore necessary to design a work-flow that makes it possible to deal with data collections, even if they are initially available as photos, scans, wikipedia dumps, etc.

These challenges are particularly relevant when dealing with ancient languages or scripts where fonts, OCR techniques, language models may be not extant or of inferior quality, as a result, among others, of the variety of writing systems and the lack of textual data. We will therefore work on improving print OCR for some of these languages, especially by moving towards joint OCR and language models. Of course, contemporary texts can be often gathered in very large volumes, as we already do within the ANR project SoSweet, resulting in different, specific issues.

ALMAnaCH pays a specific attention to the re-usability [0] of all resources produced and maintained within its various projects and research activities. To this end, we will ensure maximum compatibility with available international standards for representing textual sources and their annotations. More precisely we will take the TEI (*Text Encoding Initiative*) guidelines as well the standards produced by ISO committee TC 37/SC 4 as essential points of reference.

From our ongoing projects in the field of Digital Humanities and emerging initiatives in this field, we observe a real need for complete but easy work-flows for exploiting corpora, starting from a set of raw documents and reaching the level where one can browse the main concepts and entities, explore their relationship, extract specific pieces of information, always with the ability to return to (fragments of) the original documents. The pieces of information extracted from the corpora also need to be represented as knowledge databases (for instance as RDF "linked data"), published and linked with other existing databases (for instance for people and locations).

The process may be seen as progressively enriching the documents with new layers of annotations produced by various NLP modules and possibly validated by users, preferably in a collaborative way. It relies on the use of clearly identified representation formats for the annotations, as advocated within ISO TC 37/SC 4 standards and the TEI guidelines, but also on the existence of well-designed collaborative interfaces for browsing, querying, visualisation, and validation. ALMAnaCH has been or is working on several of the NLP bricks needed for setting such a work-flow, and has a solid expertise in the issues related to standardisation (of documents and annotations). However, putting all these elements in a unified work-flow that is simple to deploy and configure remains to be done. In particular, work-flow and interface should maybe not be dissociated, in the sense that the work-flow should be easily piloted and configured from the interface. An option will be to identify pertinent emerging platforms in DH (such as Transkribus) and to propose collaborations to ensure that NLP modules can be easily integrated.

It should be noted that such work-flows have actually a large potential besides DH, for instance for exploiting internal documentation (for a company) or exploring existing relationships between entities.

3.4.2. Development of Lexical Resources

ALPAGE, the Inria predecessor of ALMAnaCH, has put a strong emphasis in the development of morphological, syntactic and wordnet-like semantic lexical resources for French as well as other languages (see for instance [5], [1]). Such resources play a crucial role in all NLP tools, as has been proven among other tasks for POS tagging [101], [97], [111] and parsing, and some of the lexical resource development will be targeted towards the improvement of NLP tools. They will also play a central role for studying diachrony in the lexicon, for example for Ancient to Contemporary French in the context of the Profiterole project. They will also be one of the primary sources of linguistic information for augmenting language models used in OCR systems for ancient scripts, and will allow us to develop automatic annotation tools (e.g. POS taggers) for low-resourced languages (see already [113]), especially ancient languages. Finally, semantic lexicons such as wordnets will play a crucial role in assessing lexical similarity and automating etymological research.

[0] From a larger point of view we intend to comply with the so-called FAIR principles (http://force11.org/group/fairgroup/fairprinciples).

Therefore, an important effort towards the development of new morphological lexicons will be initiated, with a focus on ancient languages of interest. Following previous work by ALMAnaCH members, we will try and leverage all existing resources whenever possible such as electronic dictionaries, OCRised dictionaries, both modern and ancient [100], [83], [102], while using and developing (semi)automatic lexical information extraction techniques based on existing corpora [98], [104]. A new line of research will be to integrate the diachronic axis by linking lexicons that are in diachronic relation with one another thanks to phonetic and morphological change laws (e.g. XIIth century French with XVth century French and contemporary French). Another novelty will be the integration of etymological information in these lexical resources, which requires the formalisation, the standardisation, and the extraction of etymological information from OCRised dictionaries or other electronic resources, as well as the automatic generation of candidate etymologies. These directions of research are already investigated in ALMAnaCH [83], [102].

An underlying effort for this research will be to further the development of the GROBID-dictionaries software, which provides cascading CRF (Conditional Random Fields) models for the segmentation and analysis of existing print dictionaries. The first results we have obtained have allowed us to set up specific collaborations to improve our performances in the domains of a) recent general purpose dictionaries such as the Petit Larousse (Nénufar project, funded by the DGLFLF in collaboration with the University of Montpellier), b) etymological dictionaries (in collaboration with the Berlin Brandenburg Academy of sciences) and c) patrimonial dictionaries such as the Dictionnaire Universel de Basnage (an ANR project, including a PhD thesis at ALMAnaCH, has recently started on this topic in collaboration with the University of Grenoble-Alpes and the University Sorbonne Nouvelle in Paris).

In the same way as we signalled the importance of standards for the representation of interoperable corpora and their annotations, we will keep making the best use of the existing standardisation background for the representation of our various lexical resources. There again, the TEI guidelines play a central role, and we have recently participated in the "TEI Lex 0" initiative to provide a reference subset for the "Dictionary" chapter of the guidelines. We are also responsible, as project leader, of the edition of the new part 4 of the ISO standard 24613 (LMF, Lexical Markup Framework) [94] dedicated to the definition of the TEI serialisation of the LMF model (defined in ISO 24613 part 1 'Core model', 2 'Machine Readable Dictionaries' and 3 'Etymology'). We consider that contributing to standards allows us to stabilise our knowledge and transfer our competence.

3.4.3. Development of Annotated Corpora

Along with the creation of lexical resources, ALMAnaCH is also involved in the creation of corpora either fully manually annotated (gold standard) or automatically annotated with state-of-the-art pipeline processing chains (silver standard). Annotations will either be only morphosyntactic or will cover more complex linguistic levels (constituency and/or dependency syntax, deep syntax, maybe semantics). Former members of the ALPAGE project have a renowned experience in those aspects (see for instance [107], [93], [106], [88]) and will participate to the creation of valuable resources originating from the historical domain genre.

Under the auspices of the ANR Parsiti project, led by ALMAnaCH (PI: DS), we aim to explore the interaction of extra-linguistic context and speech acts. Exploiting extra-linguistics context highlights the benefits of expanding the scope of current NLP tools beyond unit boundaries. Such information can be of spatial and temporal nature, for instance. They have been shown to improve Entity Linking over social media streams [76]. In our case, we decided to focus on a closed world scenario in order to study context and speech acts interaction. To do so, we are developing a multimodal data set made of live sessions of a first person shooter video game (Alien vs. Predator) where we transcribed all human players interactions and face expressions streamlined with a log of all in-game events linked to the video recording of the game session, as well as the recording of the human players themselves. The in-games events are ontologically organised and enable the modelling of the extra-linguistics context with different levels of granularity. Recorded over many games sessions, we already transcribed over 2 hours of speech that will serve as a basis for exploratory work, needed for the prototyping of our context-enhanced NLP tools. In the next step of this line of work, we will focus on enriching this data set with linguistic annotations, with an emphasis on co-references resolutions and predicate

argument structures. The midterm goal is to use that data set to validate a various range of approaches when facing multimodal data in a close-world environment.

COML Team

3. Research Program

3.1. Background

In recent years, Artificial Intelligence (AI) has achieved important landmarks in matching or surpassing human level performance on a number of high level tasks (playing chess and go, driving cars, categorizing picture, etc., [31], [34], [39], [30], [36]). These strong advances were obtained by deploying on large amounts of data, massively parallel learning architectures with simple brain-inspired 'neuronal' elements. However, humans brains still outperform machines in several key areas (language, social interactions, common sense reasoning, motor skills), and are more flexible : Whereas machines require extensive expert knowledge and massive training for each particular application, humans learn autonomously over several time scales: over the developmental scale (months), humans infants acquire cognitive skills with noisy data and little or no expert feedback (weakly/unsupervised learning)[1]; over the short time scale (minutes, seconds), humans combine previously acquired skills to solve new tasks and apply rules systematically to draw inferences on the basis of extremely scarce data (learning to learn, domain adaptation, one- or zero-shot learning) [33].

The general aim of CoML, following the roadmap described in [1], is to bridge the gap in cognitive flexibility between humans and machines learning in language processing and common sense reasoning by reverse engineering how young children between 1 and 4 years of age learn from their environment. We conduct work along two axes: the first one, which we called *Developmental AI* is focused on building infant inspired machine learning algorithms. The second axis is devoted to using the developed algorithms to conduct *quantitative studies* of how infant learn across diverse environments.

3.2. Weakly/Unsupervised Learning

Much of standard machine learning is construed as regression or classification problems (mapping input data to expert-provided labels). Human infants rarely learn in this fashion, at least before going to school: they learn language, social cognition, and common sense autonomously (without expert labels) and when adults provide feedback, it is ambiguous and noisy and cannot be taken as a gold standard. Modeling or mimicking such achievement requires deploying unsupervised or weakly supervised algorithms which are less well known than their supervised counterparts.

We take inspiration from infant's landmarks during their first years of life: they are able to learn acoustic models, a lexicon, and susbtantive elements of language models and world models from raw sensory inputs. Building on previous work [3], [7], [11], we use DNN and Bayesian architectures to model the emergence of linguistic representations without supervision. Our focus is to establish how the labels in supervised settings can be replaced by weaker signals coming either from multi-modal input or from hierarchically organised linguistic levels.

At the level of phonetic representations, we study how cross-modal information (lips and self feedback from articulation) can supplement top-down lexical information in a weakly supervised setting. We use Siamese architectures or Deep CCA algorithms to combine the different views. We study how an attentional framework and uncertainty estimation can flexibly combine these informations in order to adapt to situations where one view is selectively degraded.

At the level of lexical representations, we study how audio/visual parallel information (ie. descriptions of images or activities) can help in segmenting and clustering word forms, and vice versa, help in deriving useful visual features. To achieve this, we will use architectures deployed in image captioning or sequence to sequence translation [37].

At the level of semantic and conceptual representations, we study how it is possible to learn elements of the laws of physics through the observation of videos (object permanence, solidity, spatio-temporal continuity, inertia, etc.), and how objects and relations between objects are mapped onto language.

3.3. Evaluating Machine Intelligence

Increasingly, complicated machine learning systems are being incorporated into real-life applications (e.g. self-driving cars, personal assistants), even though they cannot be formally verified, guaranteed statistically, nor even explained. In these cases, a well defined *empirical approach* to evaluation can offer interesting insights into the functioning and offer some control over these algorithms.

Several approaches exist to evaluate the 'cognitive' abilities of machines, from the subjective comparison of human and machine performance [38] to application-specific metrics (e.g., in speech, word error rate). A recent idea consist in evaluating an AI system in terms of it's *abilities* [32] , i.e., functional components within a more global cognitive architecture [35]. Psychophysical testing can offer batteries of tests using simple tasks that are easy to understand by humans or animals (e.g, judging whether two stimuli are same or different, or judging whether one stimulus is 'typical') which can be made selective to a specific component and to rare but difficult or adversarial cases. Evaluations of learning rate, domain adaptation and transfer learning are simple applications of these measures. Psychophysically inspired tests have been proposed for unsupervised speech and language learning [10], [6].

3.4. Documenting human learning

Infants learn their first language in a spontaneous fashion, across a lot of variation in amount of speech and the nature of the infant/adult interaction. In some linguistic communities, adults barely address infants until they can themselves speak. Despite these large variations in quantity and content, language learning proceeds at similar paces. Documenting such resilience is an essential step in understanding the nature of the learning algorithms used by human infants. Hence, we propose to collect and/or analyse large datasets of inputs to infants and correlate this with outcome measure (phonetic learning, vocabulary growth, syntactic learning, etc.).

MULTISPEECH Project-Team

3. Research Program

3.1. Beyond black-box supervised learning

This research axis focuses on fundamental, domain-agnostic challenges relating to deep learning, such as the integration of domain knowledge, data efficiency, or privacy preservation. The results of this axis naturally apply in the domains studied in the two other research axes.

3.1.1. Integrating domain knowledge

State-of-the-art methods in speech and audio are based on neural networks trained for the targeted task. This paradigm faces major limitations: lack of interpretability and of guarantees, large data requirements, and inability to generalize to unseen classes or tasks. We intend to research **deep generative models** as a way to learn task-agnostic probabilistic models of audio signals and design inference methods to combine and reuse them for a variety of tasks. We will pursue our investigation of hybrid methods that combine the representational power of deep learning with **statistical signal processing** expertise by leveraging recent optimization techniques for non-convex, non-linear inverse problems. We will also explore the integration of deep learning and **symbolic reasoning** to increase the generalization ability of deep models and to empower researchers/engineers to improve them.

3.1.2. Learning from little/no labeled data

While fully labeled data are costly, unlabeled data are cheap but provide intrinsically less information. **Weakly supervised learning** based on not-so-expensive incomplete and/or noisy labels is a promising middle ground. This entails modeling label noise and leveraging it for unbiased training. Models may depend on the labeler, the spoken context (voice command), or the temporal structure (ambient sound analysis). We will also keep studying **transfer learning** to adapt an expressive (audiovisual) speech synthesizer trained on a given speaker to another speaker for which only neutral voice data has been collected.

3.1.3. Preserving privacy

Some voice technology companies process users' voices in the cloud and store them for training purposes, which raises privacy concerns. We aim to **hide speaker identity** and (some) speaker states and traits from the speech signal, and evaluate the resulting automatic speech/speaker recognition accuracy and subjective quality/intelligibility/identifiability, possibly after removing private words from the training data. We will also explore **semi-decentralized learning** methods for model personalization, and seek to obtain statistical guarantees.

3.2. Speech production and perception

This research axis covers topics related to the production of speech through articulatory modeling and multimodal expressive speech synthesis, and topics related to the perception of speech through the categorization of sounds and prosody in native and in non-native speech.

3.2.1. Articulatory modeling

Articulatory speech synthesis will rely on further 2D and 3D modeling of the vocal tract as well as of the **dynamics of the vocal tract** from real-time MRI data. The prediction of glottis opening will also be considered so as to produce better quality acoustic events for consonants. The **coarticulation model** developed to handle the animation of the visible articulators will be extended to control the face and the tongue. This will help characterize links between the vocal tract and the face, and illustrate inner mouth articulation to learners. The suspension of articulatory movements in stuttering speech will also be studied.

3.2.2. Multimodal expressive speech

The dynamic realism of the animation of the talking head, which has a direct impact on audiovisual intelligibility, will continue to be our goal. Both the **animation** of the lower part of the face relating to speech and of the upper part relating to the facial expression will be considered, and development will continue towards a multilingual talking head. We will investigate further the modeling of **expressivity** both for audio-only and for audiovisual speech synthesis. We will also evaluate the benefit of the talking head in various use cases, including children with language and learning disabilities or deaf people.

3.2.3. Categorization of sounds and prosody

Reading and speaking are basic skills that need to be mastered. Further analysis of schooling experience will allow a better understanding of reading acquisition, especially for children with some language impairment. With respect to L1/L2 language interference [0] , a special focus will be set on the impact of L2 prosody on segmental realizations. Prosody will also be considered for its implication on the structuration of speech communication, including on discourse particles. Moreover, we will experiment the usage of speech technologies for computer assisted language learning in middle and high schools, and, hopefully, also for helping children learning to read.

3.3. Speech in its environment

The themes covered by this research axis correspond to the acoustic environment analysis, to speech enhancement and noise robustness, and to linguistic and semantic processing.

3.3.1. Acoustic environment analysis

Audio scene analysis is key to characterize the environment in which spoken communication may take place. We will investigate audio event detection methods that exploit both strongly/weakly labeled and unlabeled data, operate in real-world conditions, can discover novel events, and provide a semantic interpretation. We will keep working on source localization in the presence of nearby acoustic reflectors. We will also pursue our effort at the interface of **room acoustics** to blindly estimate room properties and develop acoustics-aware signal processing methods. Beyond spoken communication, this has many applications to surveillance, robot audition, building acoustics, and augmented reality.

3.3.2. Speech enhancement and noise robustness

We will pursue **speech enhancement** methods targeting several distortions (echo, reverberation, noise, overlapping speech) for both speech and speaker recognition applications, and extend them to ad-hoc arrays made of the microphones available in our daily life using multi-view learning. We will also continue to explore statistical signal models **beyond the usual zero-mean complex Gaussian model** in the time-frequency domain, e.g., deep generative models of the signal phase. **Robust acoustic modeling** will be achieved by learning domain-invariant representations or performing unsupervised domain adaptation on the one hand, and by extending our uncertainty-aware approach to more advanced (e.g., nongaussian) uncertainty models and accounting for the additional uncertainty due to short utterances on the other hand, with application to speaker and language recognition “in the wild”.

3.3.3. Linguistic and semantic processing

We will seek to address robust speech recognition by exploiting word/sentence embeddings carrying **semantic information** and combining them with acoustical uncertainty to rescore the recognizer outputs. We will also combine semantic content analysis with text obfuscation models (similar to the label noise models to be investigated for weakly supervised training of speech recognition) for the task of detecting and classifying (hateful, aggressive, insulting, ironic, neutral, etc.) **hate speech** in social media.

[0] L1 refers to the speaker’s native language, and L2 to a speaker’s second language, usually learned later as a foreign language

PANAMA Project-Team

3. Research Program

3.1. Axis 1: Sparse Models and Representations

3.1.1. Efficient Sparse Models and Dictionary Design for Large-scale Data

Sparse models are at the core of many research domains where the large amount and high-dimensionality of digital data requires concise data descriptions for efficient information processing. Recent breakthroughs have demonstrated the ability of these models to provide concise descriptions of complex data collections, together with algorithms of provable performance and bounded complexity.

A crucial prerequisite for the success of today's methods is the knowledge of a "dictionary" characterizing how to concisely describe the data of interest. Choosing a dictionary is currently something of an "art", relying on expert knowledge and heuristics.

Pre-chosen dictionaries such as wavelets, curvelets or Gabor dictionaries, are based upon stylized signal models and benefit from fast transform algorithms, but they fail to fully describe the content of natural signals and their variability. They do not address the huge diversity underlying modern data much beyond time series and images: data defined on graphs (social networks, internet routing, brain connectivity), vector valued data (diffusion tensor imaging of the brain), multichannel or multi-stream data (audiovisual streams, surveillance networks, multimodal biomedical monitoring).

The alternative to a pre-chosen dictionary is a trained dictionary learned from signal instances. While such representations exhibit good performance on small-scale problems, they are currently limited to low-dimensional signal processing due to the necessary training data, memory requirements and computational complexity. Whether designed or learned from a training corpus, dictionary-based sparse models and the associated methodology fail to scale up to the volume and resolution of modern digital data, for they intrinsically involve difficult linear inverse problems. To overcome this bottleneck, a new generation of efficient sparse models is needed, beyond dictionaries, encompassing the ability to provide sparse and structured data representations as well as computational efficiency. For example, while dictionaries describe low-dimensional signal models in terms of their "synthesis" using few elementary building blocks called atoms, in "analysis" alternatives the low-dimensional structure of the signal is rather "carved out" by a set of equations satisfied by the signal. Linear as well as nonlinear models can be envisioned.

3.1.2. Compressive Learning

A flagship emerging application of sparsity is the paradigm of compressive sensing, which exploits sparse models at the analog and digital levels for the acquisition, compression and transmission of data using limited resources (fewer/less expensive sensors, limited energy consumption and transmission bandwidth, etc.). Besides sparsity, a key pillar of compressive sensing is the use of random low-dimensional projections. Through compressive sensing, random projections have shown their potential to allow drastic dimension reduction with controlled information loss, provided that the projected signal vector admits a sparse representation in some transformed domain. A related scientific domain, where sparsity has been recognized as a key enabling factor, is Machine Learning, where the overall goal is to design statistically founded principles and efficient algorithms in order to infer general properties of large data collections through the observation of a limited number of representative examples. Marrying sparsity and random low-dimensional projections with machine learning shall allow the development of techniques able to efficiently capture and process the information content of large data collections. The expected outcome is a dramatic increase of the impact of sparse models in machine learning, as well as an integrated framework from the signal level (signals and their acquisition) to the semantic level (information and its manipulation), and applications to data sizes and volumes of collections that cannot be handled by current technologies.

3.2. Axis 2: Robust Acoustic Scene Analysis

3.2.1. Compressive Acquisition and Processing of Acoustic Scenes

Acoustic imaging and scene analysis involve acquiring the information content from acoustic fields with a limited number of acoustic sensors. A full 3D+t field at CD quality and Nyquist spatial sampling represents roughly 10^6 microphones/m^3. Dealing with such high-dimensional data requires to drastically reduce the data flow by positioning appropriate sensors, and selecting from all spatial locations the few spots where acoustic sources are active. The main goal is to develop a theoretical and practical understanding of the conditions under which compressive acoustic sensing is both feasible and robust to inaccurate modeling, noisy measures, and partially failing or uncalibrated sensing devices, in various acoustic sensing scenarii. This requires the development of adequate algorithmic tools, numerical simulations, and experimental data in simple settings where hardware prototypes can be implemented.

3.2.2. Robust Audio Source Separation

Audio signal separation consists in extracting the individual sound of different instruments or speakers that were mixed on a recording. It is now successfully addressed in the academic setting of linear instantaneous mixtures. Yet, real-life recordings, generally associated to reverberant environments, remain an unsolved difficult challenge, especially with many sources and few audio channels. Much of the difficulty comes from the combination of (i) complex source characteristics, (ii) sophisticated underlying mixing model and (iii) adverse recording environments. Moreover, as opposed to the "academic" blind source separation task, most applicative contexts and new interaction paradigms offer a variety of situations in which prior knowledge and adequate interfaces enable the design and the use of informed and/or manually assisted source separation methods.

One of the objectives of PANAMA is to instantiate and validate specific instances of audio source separation approaches and to target them to real-world industrial applications, such as 5.1 movie re-mastering, interactive music soloist control and outdoor speech enhancement. Extensions of the framework are needed to achieve real-time online processing, and advanced constraints or probabilistic priors for the sources at hand need to be designed, while paying attention to computational scalability issues.

In parallel to these efforts, expected progress in sparse modeling for inverse problems shall bring new approaches to source separation and modeling, as well as to source localization, which is often an important first step in a source separation workflow.

3.2.3. Robust Audio Source Localization

Audio source localization consists in estimating the position of one or several sound sources given the signals received by a microphone array. Knowing the geometry of an audio scene is often a pre-requisite to perform higher-level tasks such as speaker identification and tracking, speech enhancement and recognition or audio source separation. It can be decomposed into two sub-tasks : (i) compute spatial auditory features from raw audio input and (ii) map these features to the desired spatial information. Robustly addressing both these aspects with a limited number of microphones, in the presence of noise, reverberation, multiple and possibly moving sources remains a key challenge in audio signal processing. The first aspect will be tackled by both advanced statistical and acoustical modeling of spatial auditory features. The second one will be addressed by two complementary approaches. *Physics-driven* approaches cast sound source localization as an inverse problem given the known physics of sound propagation within the considered system. *Data-driven* approaches aim at learning the desired feature-to-source-position mapping using real-world or synthetic training datasets adapted to the problem at hand. Combining these approaches should allow a widening of the notion of source localization, considering problems such as the identification of the directivity or diffuseness of the source as well as some of the boundary conditions of the room. A general perspective is to investigate the relations between the physical structure of the source and the particular structures that can be discovered or enforced in the representations and models used for characterization, localization and separation.

3.3. Axis 3: Large-scale Audio Content Processing and Self-organization

3.3.1. Motif Discovery in Audio Data

Facing the ever-growing quantity of multimedia content, the topic of motif discovery and mining has become an emerging trend in multimedia data processing with the ultimate goal of developing weakly supervised paradigms for content-based analysis and indexing. In this context, speech, audio and music content, offers a particularly relevant information stream from which meaningful information can be extracted to create some form of "audio icons" (key-sounds, jingles, recurrent locutions, musical choruses, etc ...) without resorting to comprehensive inventories of expected patterns.

This challenge raises several fundamental questions that will be among our core preoccupations over the next few years. The first question is the deployment of motif discovery on a large scale, a task that requires extending audio motif discovery approaches to incorporate efficient time series pattern matching methods (fingerprinting, similarity search indexing algorithms, stochastic modeling, etc.). The second question is that of the use and interpretation of the motifs discovered. Linking motif discovery and symbolic learning techniques, exploiting motif discovery in machine learning are key research directions to enable the interpretation of recurring motifs.

On the application side, several use cases can be envisioned which will benefit from motif discovery deployed on a large scale. For example, in spoken content, word-like repeating fragments can be used for several spoken document-processing tasks such as language-independent topic segmentation or summarization. Recurring motifs can also be used for audio summarization of audio content. More fundamentally, motif discovery paves the way for a shift from supervised learning approaches for content description to unsupervised paradigms where concepts emerge from the data.

3.3.2. Structure Modeling and Inference in Audio and Musical Contents

Structuring information is a key step for the efficient description and learning of all types of contents, and in particular audio and musical contents. Indeed, structure modeling and inference can be understood as the task of detecting dependencies (and thus establishing relationships) between different fragments, parts or sections of information content.

A stake of structure modeling is to enable more robust descriptions of the properties of the content and better model generalization abilities that can be inferred from a particular content, for instance via cache models, trigger models or more general graphical models designed to render the information gained from structural inference. Moreover, the structure itself can become a robust descriptor of the content, which is likely to be more resistant than surface information to a number of operations such as transmission, transduction, copyright infringement or illegal use.

In this context, information theory concepts need to be investigated to provide criteria and paradigms for detecting and modeling structural properties of audio contents, covering potentially a wide range of application domains in speech content mining, music modeling or audio scene monitoring.

SEMAGRAMME Project-Team

3. Research Program

3.1. Overview

The research program of Sémagramme aims to develop models based on well-established mathematics. We seek two main advantages from this approach. On the one hand, by relying on mature theories, we have at our disposal sets of mathematical tools that we can use to study our models. On the other hand, developing various models on a common mathematical background will make them easier to integrate, and will ease the search for unifying principles.

The main mathematical domains on which we rely are formal language theory, symbolic logic, and type theory.

3.2. Formal Language Theory

Formal language theory studies the purely syntactic and combinatorial aspects of languages, seen as sets of strings (or possibly trees or graphs). Formal language theory has been especially fruitful for the development of parsing algorithms for context-free languages. We use it, in a similar way, to develop parsing algorithms for formalisms that go beyond context-freeness. Language theory also appears to be very useful in formally studying the expressive power and the complexity of the models we develop.

3.3. Symbolic Logic

Symbolic logic (and, more particularly, proof-theory) is concerned with the study of the expressive and deductive power of formal systems. In a rule-based approach to computational linguistics, the use of symbolic logic is ubiquitous. As we previously said, at the level of syntax, several kinds of grammars (generative, categorial...) may be seen as basic deductive systems. At the level of semantics, the meaning of an utterance is captured by computing (intermediate) semantic representations that are expressed as logical forms. Finally, using symbolic logics allows one to formalize notions of inference and entailment that are needed at the level of pragmatics.

3.4. Type Theory and Typed λ-Calculus

Among the various possible logics that may be used, Church's simply typed λ-calculus and simple theory of types (a.k.a. higher-order logic) play a central part. On the one hand, Montague semantics is based on the simply typed λ-calculus, and so is our syntax-semantics interface model. On the other hand, as shown by Gallin, the target logic used by Montague for expressing meanings (i.e., his intensional logic) is essentially a variant of higher-order logic featuring three atomic types (the third atomic type standing for the set of possible worlds).

Auctus Team

3. Research Program

3.1. Analysis and modelling of human behavior

3.1.1. Scientific Context

The purpose of this axis is to provide metrics to assess human behavior. We place ourselves here from the point of view of the human being and more precisely of the industrial operator. We assume the following working hypotheses: the operator's task and environmental conditions are known and circumscribed; the operator is trained in the task, production tools and safety instructions; the task is repeated with more or less frequent intervals. We focus our proposals on assessing:

- the physical and cognitive fragility of operators in order to meet assistance needs;
- cognitive biases and physical constraints leading to a loss of operator safety;
- ergonomic, performance and acceptance of the production tool.

In the industrial context, the fields that best answer these questions are work ergonomics and cognitive sciences. Scientists typically work on 4 axes: physiological/biomechanical, cognitive, psychological and sociological. More specifically, we focus on biomechanical, cognitive and psychological aspects, as described by the ANACT [24], [26]. The aim here is to translate these factors into metrics, optimality criteria or constraints in order to implement them in our methodologies for analysis, design and control of the collaborative robot.

To understand our desired contributions in robotics, we must review the current state of ergonomic workstation evaluation, particularly at the biomechanical level. The ergonomist evaluates the gesture through the observation of workstations and, generally, through questionnaires. This requires long periods of field observation, followed by analyses based on ergonomic grids (e.g. RULA [42], REBA [32], LUBA [37], OWAS [36], ROSA [60],...). Until then, the use of more complex measurement systems was reserved for laboratories, particularly biomechanical laboratories. The appearance of inexpensive sensors such as IMUs (Inertial Measurement Units) or RGB-D cameras makes it possible to consider a digitalized, and therefore objective, observation of the gesture, postures and more generally of human movement. Thanks to these sensors, which are more or less intrusive, it is now possible to permanently install observation systems on production lines. This completely changes paradigms and opens the door to longitudinal observations. It should be noted that this is comparable to the evolution of maintenance, which becomes predictive.

On the strength of this new paradigm, *ergonomic robotics* has recently taken an interest in this type of evaluation to adapt the robot's movements in order to reduce ergonomic risk scores. This approach complements the more traditional approaches that only consider the performance of the action produced by the human in interaction with the robot. However, we must go further. Indeed, the ergonomic criteria are based on the principle that the comfort positions are distant from the human articular stops. In addition, the notation must be compatible with an observation of the human being through the eye of the ergonomist. In practice, evaluations are inaccurate and subjective [63]. Moreover, they are made for quasi-static human positions without taking into account the evolution of the person's physical, physiological and psychological state. The repetition of gestures, the solicitation of muscles and joints is one of the questions that must complete these analyses. One of the methods used by ergonomists to limit biomechanical exposures is to increase variations in motor stress by rotating tasks [61]. However, this type of extrinsic method is not always possible in the industrial context [40].

One of Auctus' objectives is to show how, through a cobot, the operator's environment can be varied to encourage more appropriate motor strategies. To do so, we must focus on a field of biomechanics that studies the intrinsic variability of the motor system allowed by the joint redundancy of the human body. This motor variability refers to the natural alternation of postures, movements and muscle activity observed in the individual to respond to a requested task [61]. This natural variation leads to differences between the motor coordinates used by individuals, which evokes the notion of motor strategy [33].

As shown by the cognitive dimension of ergonomics (see above), we believe that some of these motor strategies are a physically quantifiable reflection of the operator's cognitive state. For example, fatigue [57] and its anticipation or the manual expertise (dexterous and cognitive) of the operator which allows him to anticipate his movements over long periods of time in order to preserve his body, his performance and his pain.

3.1.2. Methodology

How can we observe, understand and quantify these human motor strategies to better design and control the behavior of the cobotic assistant? When we study the systems of equations considered (kinematic, static, dynamic, musculoskeletal), several problems appear and explain our methodological choices:

- the large dimensions of the problems to be considered, due to joint, muscle and placement redundancy,
- the variabilities of the parameters, for example: physiological (consider not an operator, but a set of operators), geometric (consider a set of possible placements of the operator) and static (consider a set of forces that the operator must produce);
- the uncertainties of measurement, model approximation.

The idea is to start from a description of redundant workspaces (geometric, static, dynamic...). To do this, we use set theory approaches, based on interval analysis [3], [48], which allow us to respond to the uncertainties and variability issues previously mentioned. In addition, one of the advantages of these techniques is that they allow the results to be certified, which is essential to address safety issues. Some members of the team has already achieved success in mechanical design for performance certification and robot design [44]. The adaptation of these approaches allows us to obtain a mapping of ergonomic and efficient movements in which we can project the operators' motor strategies and thus define a metric quantifying the sensorimotor commands chosen with regard to the cognitive criteria studied.

It is therefore necessary to:

- propose new indices linking different types of performance (ergonomic biomechanical robotics, but also influence of fatigue, stress, level of expertise on the evolution of performance);
- divide the gesture into homogeneous phases: this process is complex and depends on the type of index used and the techniques used. We are exploring several ways: inverse optimal control, learning methods, or the use of techniques from signal processing.
- develop interval extensions of the identified indices. These indices are not necessarily the result of a direct model, and algorithms need to be developed or adapted (calculation of manipulability, UCM, etc.).
- Aggregate proposals into a dedicated interval analysis library (use of and contribution to the existing ALIAS-Inria and the open source IBEX library).

The originality and contribution of the methodology is to allow an analysis taking into account in the same model the measurement uncertainties (important for on-site use of analytical equipment), the variability of tasks and trajectories, and the physiological characteristics of the operators.

Other avenues of research are being explored, particularly around the inverse optimal control [49] which allows us to project human movement on the basis of performance indices and thus to offer a possible interpretation in the analysis of behaviors.

We also use automatic classification techniques: 1) to propose cognitive models that will be learned experimentally 2) for segmentation or motion recognition, for example by testing Reservoir Computing [34] approaches.

3.2. Operator / robot coupling

3.2.1. Scientific Context

Thanks to the progress made in recent years in the field of p-HRI (Physical Human-Robot Interaction), robotic systems are beginning to operate in the same workspace as humans, which is profoundly changing industrial

issues and allowing a wide variety of human-robot coupling solutions to be considered to perform the same task [25]. Different types of interactions exist. They can be classified in different ways: according to the degree of autonomy of the robot and its proximity to the user [31] with particularities for "wearable robots" [30], [29], or for collaborative robotics [62], or according to the role of the human being [58]. From a cognitive point of view, classifications are more concerned with autonomy, the complexity of information processing and the type of communication and representation of the human being by the robot [47], [64].

We proposed a classification of cobotic systems according to the configuration of the schema of interactions between humans, robots and the environment [45], [55].

The parameters of the coupling being numerous and complex, the determination of the most appropriate type of coupling for a type of problem is an open problem [50], [2], [46], [41]. The traditional approach consists in trying to identify and classify the various possible options and to select the one that seems most relevant with regard to the feasibility, efficiency, budget envelope and acceptability of the operator. One of the main objectives of our research project is to define a typology of cobots or cobotic systems in order to specify the methodology for developing the best solution: what are the criteria for defining the best robotic architecture, what type of coupling, what autonomy of the robot, what role for the operator, what risks for the human, what overall performance? These are the key issues that need to be addressed. To meet this methodological need, we propose an approach guided by experience on use cases obtained thanks to our industrial partners.

3.2.2. Methodology

Task analysis and human behavior modelling, discussed in the previous sections, should help to characterize the different types of coupling and interaction modalities, their advantages and disadvantages, in order to assist in the decision-making process. One of the ideas we would like to develop is to try to break down the task into a sequence of elementary gestures corresponding to simple motor actions performed in a clearly identified context and to evaluate for each of them the degree of feasibility in automatic mode or in robot assistance mode. The assessment must take into account a large number of parameters that relate to physical interactions, human-robot communication, reliability and human factors, including acceptability and impact on the valuation or devaluation of the operator's work. Concerning the evaluation of human factors, we have already begun to work on the subject within the more general framework of human systems interactions by operating Bayesian networks, drawing inspiration from the work of [28], [56].

The adoption of assessment criteria for a single domain (e. g. robotics or ergonomics) cannot guarantee that the performance of this coupling will be maximized. From design to evaluation, cross-effects must be constantly considered:

- impact of the cobot design on the user's performance: intuitiveness, adaptation to intra- and inter-individual variations, affordance, stress factors (noise, vibrations,...), fatigue factors (control laws, necessary attention,...) and motivation factors (effectiveness, efficiency, aesthetics,...);
- impact of user performance on cobot exploitation: risks of human error (attention error, perseveration, circumvention of procedures, syndrome outside the loop) [28].

In addition to purely physical assistance, some cobotic systems are designed to assist the operator in his decision-making. The issues of trust, acceptance, sharing of representations and co-construction of a shared awareness of the situation are then to be addressed [59].

3.3. Design of cobotic systems

3.3.1. Architectural design

Is it necessary to cobotize, robotize or assist the human being? Which mechanical architecture meets the task challenges (a serial cobot, a specific mechanism, an exoskeleton)? What type of interaction (H/R cohabitation, comanipulation, teleoperation)? These questions are the first requests from our industrial partners. For the moment, we have few comprehensive methodological answers to provide them. Choosing a collaborative robot architecture is a difficult problem [38]. It is all the more when the questions are approached from both a cognitive ergonomics and robotics perspective. There are indeed major methodological and conceptual

differences in these areas. It is therefore necessary to bridge these representational gaps and to propose an approach that takes into consideration the expectations of the robotician to model and formalize the general properties of a cobotic system as well as those of the ergonomist to define the expectations in terms of an assistance tool.

To do this, we propose a user-centered design approach, with a particular focus on human-system interactions. From a methodological point of view, this requires first of all the development of a structured experimental approach aimed at characterizing the task to be carried out through a "system" analysis but also at capturing the physical markers of its realization: movements and efforts required, ergonomic stress. This characterization must be done through the prism of the systematic study of the exchange of information (and their nature) by humans in their performance of the considered task. On the basis of these analyses, the main challenge is to define a decision support tool for the choice of the robotic architecture and for the specifications of the role assigned to the robot and the operator as well as their interactions.

The evolution of the chosen methodology is for the moment empirical, based on the user cases regularly treated in the team (see sections on contracts and partnerships).

It can be summarized for the moment as:

- identify difficult jobs on industrial sites. This is done through visits and exchanges with our partners (manager, production manager, ergonomist...);
- select some of them, then observe the human in its ecological environment. Our tools allow us to produce a motion analysis, currently based on ergonomic criteria. In parallel we carry out a physical evaluation of the task in terms of expected performance and an evaluation of the operator by means of questionnaires.
- Synthesize these first results to deduce the robotic architectures to be initiated, the key points of human-robot interaction to be developed, the difficulties in terms of human factors to be taken into account.

In addition, the different human and task analyses take advantage of the different expertise available within the team. We would like to gradually introduce the evaluation criteria presented above. Indeed, the team has already worked on the current dominant approach: the use of a virtual human to design the cobotic cell through virtual tools [1]. However, the very large dimensions of the problems treated (modelling of the body's ddl and the constraints applied to it) makes it difficult to carry out a certified analysis. We then choose to go through the calculation of the body's workspace, representing its different performances, which is not yet done in this field. The idea here is to apply set theory approaches, using interval analysis and already discussed in section 3.1.2 . The goal is then to extend to intervals the constraints played in virtual reality during the simulation. This would allow the operator to check his trajectories and scenarios not only for a single case study but also for sets of cases. For example, it can be verified that, regardless of the bounded sets of simulated operator physiologies, the physical constraints of a simulated trajectory are not violated. Thus, the assisted design tools certify cases of use as a whole. Moreover, the intersection between the human and robot workspaces provides the necessary constraints to certify the feasibility of a task. This allows us to better design a cobotic system to integrate physical constraints. In the same way, we will look for ways in which human cognitive markers can be included in this approach.

Thus, we summarize here the contributions of the other research axes, from the analysis of human behavior in its environment for an identified task, to the choice of a mechanical architecture, via an evaluation of torque and interactions. All the previous analyses provide design constraints. This methodological approach is perfectly integrated into an Appropriate Design approach used for the dimensional design of robots, again based on interval analysis. Indeed, to the desired performance of the human-robot couple in relation to a task, it is sufficient to add the constraints limiting the difficulty of the operator's gesture as described above. The challenges are then the change of scale in models that symbiotically consider the human-robot pair, the uncertain, flexible and uncontrollable nature of human behavior and the many evaluation indices needed to describe them.

3.3.2. Control design

The control laws of collaborative robots from the major robot manufacturers differ little or not at all from the existing control laws in the field of conventional industrial robotics. Security is managed a posteriori, as an exception, by a security PLC / PC. It is therefore not an intrinsic property of the controller. This quite strongly restricts the possibilities of physical interaction [0] and collaboration and leads to sub-optimal operation of the robotic system. It is difficult in this context to envision real human-robot collaboration. Collaborative operation requires, in this case, a control calculation that integrates safety and ergonomics as a priori constraints.

The control of truly collaborative robots in an industrial context is, from our point of view, underpinned by two main issues. The first is related to the macroscopic adaptation of the robot's behaviour according to the phases of the production process. The second is related to the fine adaptation of the degree and/or nature of the robot's assistance according to the ergonomic state of the operator. If this second problem is part of a historical dynamic in robotics that consists in placing safety constraints, particularly those related to the presence of a human being, at the heart of the control problem [31] [43], [35], it is not approached from the more subtle point of view of ergonomics where the objective cannot be translated only in terms of human life or death, but rather in terms of long-term respect for their physical and mental integrity. Thus, the simple and progressive adoption by a human operator of the collaborative robot intended to assist him in his gesture requires a self-adaptation in the time of the command. This self-adaptation is a fairly new subject in the literature [51], [52].

At the macroscopic level, the task plan to be performed for a given industrial operation can be represented by a finite state machine. In order not to increase the human's cognitive load by explicitly asking him to manage transitions for the robot, we propose to develop a decision algorithm to ensure discrete transitions from one task (and the associated assistance mode) to another based on an online estimate of the current state of the human-robot couple. The associated scientific challenge requires establishing a link between the robot's involvement and a given working situation. To do so, we propose an incremental approach to learning this complex relationship. The first stage of this work will consist in identifying the general and relevant control variables to conduct this learning in an efficient and reusable way, regardless of the particular method of calculating the control action. Physically realistic simulations and real word experiments will be used to feed this learning process.
In order to handle mode transitions, we propose to explore the richness of the multi-tasking control formalism under constraints [39] in order to ensure a continuous transition from one control mode to another while guaranteeing compliance with a certain number of control constraints. Some of these constraints are based on ergonomic specifications and are dependent on the state of the robot and of the human operator, which, by nature, is difficult to predict accurately. We propose to exploit the interval analysis paradigm to efficiently formulate ergonomic constraints robust to the various existing uncertainties.
Purely discrete or reactive adaptation of the control law would make no sense given the slow dynamics of certain physiological phenomena such as fatigue. Thus, we propose to formulate the control problem as a predictive problem where the impact of the control decision at a time t is anticipated at different time horizons. This requires a prediction of human movement and knowledge of the motor variability strategies it employs. This prediction is possible on the basis of the supervision at all times of the operational objectives (task in progress) in the short term. However, it requires the use of a virtual human model and possibly a dynamic simulation to quantify the impact of these potential movements in terms of performance, including ergonomics. It is impractical to use a predictive command with simulation in the loop with an advanced virtual manikin model. We therefore propose to adapt the prediction horizon and the complexity of the corresponding model in order to guarantee a reasonable computational complexity.
The planned developments require both an approach to modelling human sensorimotor behaviour, particularly in terms of accommodating fatigue via motor variability, and validating related models in an experimental framework based on observation of movement and quantification of ergonomic performance. Experimental developments must also focus on the validation of proposed control approaches in concrete contexts. To begin with, the Woobot project related to gesture assistance for carpenters (Nassim Benhahib's thesis) and

[0] In the ISO TS 15066 technical specification on collaborative robotics, human-robot physical interaction is allowed but perceived as a situation to be avoided.

a collaboration currently being set up with Safran on assistance to operators in shrink-wrapping tasks (manual knotting) in aeronautics are rich enough background elements to support the research conducted. Collaborative research projects with PSA will also soon provide a larger set of contexts in which the proposed research can be validated.

CHORALE Team

3. Research Program

3.1. Task based world modeling and understanding

Executing a robotic task needs to specify a task space and a set of objective functions to be optimized. One research issue will be to define a framework allowing to represent the tasks in a generic canonic space in order to make their design and their analysis easier thanks to the control theory tools (observability, controllability, robustness...). All along the execution of the task, autonomous robotics systems have to acquire and maintain a model of the world and of the interactions between the different components involved in the task (heterogeneous robots, human beings, changes in the environment...). This model evolves in time and in space. In this research axis, we will investigate novel task-oriented world multi-layers representations (photometry, geometry, semantic) embedded in a short/long term memory framework able to handle static and dynamic events (long term mapping). A particular attention will be also paid to integrate human-robot interactions in shared environment (social skills). Another ambition of the project will be to build a bridge between model-based and machine learning methods. Understanding the world evolution is one of the key of autonomy. In this aim, we will focus on situation awareness.

3.2. Multi-sensory perception and control

Multi-sensory based perception and control is an area that starts from one single robot evolving in the environment with a set of sensors, up to a set of heterogeneous robots collaborating for the execution of a global shared task. We will address problems such as the active selection of the most suitable source of information (e.g. sensors and features) during the execution of the task and the active sensing control in order to maximize the collected information about the world modeling (including calibration and environment parameters, exogenous disturbances), allowing the task-driven sensor-based control framework to be more efficiently and robustly executed. Another issue will be the execution of a task defined by another robot or human, and to be replicated with a robot with different capabilities in perception, control and level of autonomy (i.e. heterogeneous robots). Last issues will come from the collaboration of different autonomous and heterogeneous robots in order to accomplish a shared task (mapping, robust localization, calibration, tracking, transporting, moving, ...)

CHROMA Project-Team

3. Research Program

3.1. Introduction

The Chroma team aims to deal with different issues of autonomous mobile robotics : perception, decision-making and cooperation. Figure 1 schemes the different themes and sub-themes investigated by Chroma.

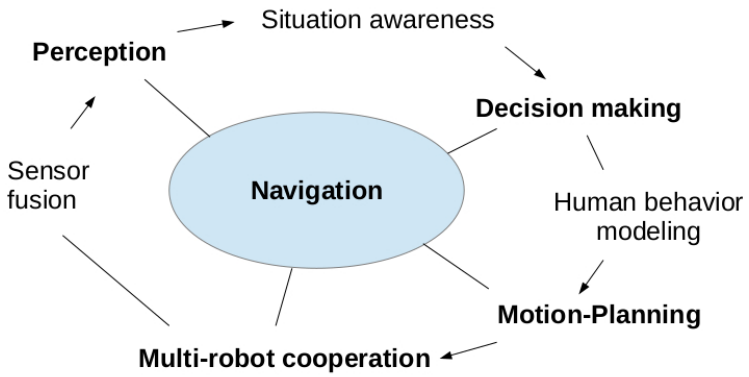


Figure 1. Research themes of the team and their relation

We present here after our approaches to address these different themes of research, and how they combine altogether to contribute to the general problem of robot navigation. Chroma pays particular attention to the problem of autonomous navigation in highly dynamic environments populated by humans and cooperation in multi-robot systems. We share this goal with other major robotic laboratories/teams in the world, such as Autonomous Systems Lab at ETH Zurich, Robotic Embedded Systems Laboratory at USC, KIT [0] (Prof Christoph Stiller lab and Prof Ruediger Dillmann lab), UC Berkeley, Vislab Parma (Prof. Alberto Broggi), and iCeiRA [0] laboratory in Taipei, to cite a few. Chroma collaborates at various levels (visits, postdocs, research projects, common publications, etc.) with most of these laboratories, see Section 9.3 .

3.2. Perception and Situation Awareness

Robust perception in open and dynamic environments populated by human beings is an open and challenging scientific problem. Traditional perception techniques do not provide an adequate solution for these problems, mainly because such environments are uncontrolled [0] and exhibit strong constraints to be satisfied (in particular high dynamicity and uncertainty). This means that **the proposed solutions have to simultaneously take into account characteristics such as real time processing, temporary occultations, dynamic changes or motion predictions**.

[0] Karlsruhe Institut fur Technologie
[0] International Center of Excellence in Intelligent Robotics and Automation Research.
[0] partially unknown and open

3.2.1. Bayesian perception

Context. Perception is known to be one of the main bottlenecks for robot motion autonomy, in particular when navigating in open and dynamic environments is subject to strong real-time and uncertainty constraints. In order to overcome this difficulty, we have proposed in the scope of the former e-Motion team, a new paradigm in robotics called "Bayesian Perception". The foundation of this approach relies on the concept of "Bayesian Occupancy Filter (BOF)" initially proposed in the Ph.D. thesis of Christophe Coue [65] and further developed in the team [0]. The basic idea is to combine a Bayesian filter with a probabilistic grid representation of both the space and the motions. It allows the filtering and the fusion of heterogeneous and uncertain sensors data, by taking into account the history of the sensors measurements, a probabilistic model of the sensors and of the uncertainty, and a dynamic model of the observed objects motions.

In the scope of the Chroma team and of several academic and industrial projects (in particular the IRT Security for autonomous vehicle and Toyota projects), we went on with the development and the extension under strong embedded implementation constraints, of our Bayesian Perception concept. This work has already led to the development of more powerful models and more efficient implementations, e.g. the *CMCDOT* (Conditional Monte Carlo Dense Occupancy Tracker) framework [89] which is still under development.

This work is currently mainly performed in the scope of the "Security for Autonomous Vehicle (SAV)" project (IRT Nanoelec), and more recently in cooperation with some Industrial Companies (see section New Results for more details on the non confidential industrial cooperation projects).

Objectives. We aim at defining a complete framework extending the Bayesian Perception paradigm to the object level. The main objective is to be simultaneously more robust, more efficient for embedded implementations, and more informative for the subsequent scene interpretation step (Figure 2 .a illustrates). Another objective is to improve the efficiency of the approach (by exploiting the highly parallel characteristic of our approach), while drastically reducing important factors such as the required memory size, the size of the hardware component, its price and the required energy consumption. This work is absolutely necessary for studying embedded solutions for the future generation of mobile robots and autonomous vehicles. We also aim at developing strong partnerships with non-academic partners in order to adapt and move the technology closer to the market.

Figure 2. a. Illustration of the Bayesian Perception Paradigm: Filtered occupancy grids, enhanced with motion estimations (vectors) and object detection (colored boxes) b. Autonomous Zoe car of Inria/Chroma.

[0]The Bayesian programming formalism developed in e-Motion, pioneered (together with the contemporary work of Thrun, Burgard and Fox [96]) a systematic effort to formalize robotics problems under Probability theory–an approach that is now pervasive in Robotics.

3.2.2. System validation

Context. Testing and validating Cyber Physical Systems which are designed for operating in various real world conditions, is both an open scientific question and a necessity for a future deployment of such systems. In particular, this is the case for Embedded Perception and Decision-making Systems which are designed for future ADAS [0] and Autonomous Vehicles. Indeed, it is unrealistic to try to be exhaustive by making a huge number of experiments in various real situations. Moreover, such experiments might be dangerous, highly time consuming, and expensive. This is why we have decided to develop appropriate *realistic simulation and statistical analysis tools* for being able to perform a huge number of tests based on some previously recorded real data and on random changes of some selected parameters (the "co-simulation" concept). Such an approach might also be used in a training step of a machine learning process. This is why simulation-based validation is getting more and more popular in automotive industry and research.

This work is performed in the scope of both the SAV [0] project (IRT Nanoelec) and of the EU Enable-S3 project; it is also performed in cooperation with the Inria team Tamis in Rennes, with the objective to integrate the Tamis "Statistical Model Checking" (SMC) approach into our validation process. We are also starting to work on this topic with the Inria team Convecs, with the objective to also integrate formal methods into our validation process.

Objectives. We started to work on this new research topic in 2017. The first objective is to build a "simulated navigation framework" for: (1) constructing realistic testing environments (including the possibility of using real experiments records), (2) developing for each vehicle a simulation model including various physical and dynamic characteristics (e.g. physics, sensors and motion control), and (3) evaluating the performances of a simulation run using appropriate statistical software tools.

The second objective is to develop models and tools for automating the Simulation & Validation process, by using a selection of relevant randomized parameters for generating large database of tests and statistical results. Then, a metric based on the use of some carefully selected "Key Performance Indicator" (KPI) has to be defined for performing a statistical evaluation of the results (e.g. by using the above-mentioned SMC approach).

3.2.3. Situation Awareness and Prediction

Context. Predicting the evolution of the perceived moving agents in a dynamic and uncertain environment is mandatory for being able to safely navigate in such an environment. We have recently shown that an interesting property of the Bayesian Perception approach is to generate short-term conservative [0] predictions on the likely future evolution of the observed scene, even if the sensing information is temporary incomplete or not available [84]. But in human populated environments, estimating more abstract properties (e.g. object classes, affordances, agent's intentions) is also crucial to understand the future evolution of the scene. This work is carried out in the scope of the Security of Autonomous Vehicle (SAV) project (IRT Nanoelec) and of several cooperative and PhD projects with Toyota and with Renault.

Objectives. The first objective is to develop an integrated approach for "Situation Awareness & Risk Assessment" in complex dynamic scenes involving multiples moving agents (e.g vehicles, cyclists, pedestrians ...), whose behaviors are most of the time unknown but predictable. Our approach relies on combining machine learning to build a model of the agent behaviors and generic motion prediction techniques (e.g. Kalman-based, GHMM , or Gaussian Processes). In the perspective of a long-term prediction we will consider the semantic level [0] combined with planning techniques.

[0] Advance Driving Assistance System
[0] Security for Autonomous Vehicles
[0] i.e. when motion parameters are supposed to be stable during a small amount of time
[0] knowledge about agent's activities and tasks

The second objective is to build a general framework for perception and decision-making in multi-robot/vehicle environments. The navigation will be performed under both dynamic and uncertainty constraints, with contextual information and a continuous analysis of the evolution of the probabilistic collision risk. Interesting published and patented results [76] have already been obtained in cooperation with Renault and UC Berkeley, by using the "Intention / Expectation" paradigm and Dynamic Bayesian Networks. We are currently working on the generalization of this approach, in order to take into account the dynamics of the vehicles and multiple traffic participants. The objective is to design a new framework, allowing us to overcome the shortcomings of rules-based reasoning approaches which often show good results in low complexity situations, but which lead to a lack of scalability and of long terms predictions capabilities.

3.2.4. Robust state estimation (Sensor fusion)

Context. In order to safely and autonomously navigate in an unknown environment, a mobile robot is required to estimate in real time several physical quantities (e.g., position, orientation, speed). These physical quantities are often included in a common state vector and their simultaneous estimation is usually achieved by fusing the information coming from several sensors (e.g., camera, laser range finder, inertial sensors). The problem of fusing the information coming from different sensors is known as the *Sensor Fusion* problem and it is a fundamental problem which plays a major role in robotics.

Objective. A fundamental issue to be investigated in any sensor fusion problem is to understand whether the state is observable or not. Roughly speaking, we need to understand if the information contained in the measurements provided by all the sensors allows us to carry out the estimation of the state. If the state is not observable, we need to detect a new observable state. This is a fundamental step in order to properly define the state to be estimated. To achieve this goal, we apply standard analytic tools developed in control theory together with some new theoretical concepts we introduced in [78] (concept of continuous symmetry). Additionally, we want to account the presence of disturbances in the observability analysis.

Our approach is to introduce general analytic tools able to derive the observability properties in the nonlinear case when some of the system inputs are unknown (and act as disturbances). We recently obtained a simple analytic tool able to account the presence of unknown inputs [80], which extends a heuristic solution derived by the team of Prof. Antonio Bicchi [60] with whom we collaborate (Centro Piaggio at the University of Pisa).

Fusing visual and inertial data. A special attention is devoted to the fusion of inertial and monocular vision sensors (which have strong application for instance in UAV navigation). The problem of fusing visual and inertial data has been extensively investigated in the past. However, most of the proposed methods require a state initialization. Because of the system nonlinearities, lack of precise initialization can irreparably damage the entire estimation process. In literature, this initialization is often guessed or assumed to be known [70]. Recently, this sensor fusion problem has been successfully addressed by enforcing observability constraints [74] and by using optimization-based approaches [77]. These optimization methods outperform filter-based algorithms in terms of accuracy due to their capability of relinearizing past states. On the other hand, the optimization process can be affected by the presence of local minima. We are therefore interested in a deterministic solution that analytically expresses the state in terms of the measurements provided by the sensors during a short time-interval.

For some years we explore deterministic solutions as presented in [79] and [81]. Our objective is to improve the approach by taking into account the biases that affect low-cost inertial sensors (both gyroscopes and accelerometers) and to exploit the power of this solution for real applications. This work is currently supported by the ANR project VIMAD [0] and experimented with a quadrotor UAV. We have a collaboration with Prof. Stergios Roumeliotis (the leader of the MARS lab at the University of Minnesota) and with Prof. Anastasios Mourikis from the University of California Riverside. Regarding the usage of our solution for real applications we have a collaboration with Prof. Davide Scaramuzza (the leader of the Robotics and Perception group at the University of Zurich) and with Prof. Roland Siegwart from the ETHZ.

[0] Navigation autonome des drones aériens avec la fusion des données visuelles et inertielles, lead by A. Martinelli, Chroma.

3.3. Navigation and cooperation in dynamic environments

In his reference book *Planning algorithms*[0] S. LaValle discusses the different dimensions that made the motion-planning problem complex, which are the number of robots, the obstacle region, the uncertainty of perception and action, and the allowable velocities. In particular, it is emphasized that complete algorithms require at least exponential time to deal with multiple robot planning in complex environments, preventing them to be scalable in practice. Moreover, dynamic and uncertain environments, as human-populated ones, expand this complexity.

In this context, we aim at **scale up decision-making in human-populated environments and in multi-robot systems, while dealing with the intrinsic limits of the robots (computation capacity, limited communication)**.

3.3.1. Motion-planning in human-populated environment

Context. Motion planning in dynamic and human-populated environments is a current challenge of robotics. Many research teams work on this topic. We can cite the Institut of robotic in Barcelone [69], the MIT [57], the Autonomous Intelligent Systems lab in Freiburg [61], or the LAAS [85]. In Chroma, we explore different issues : **integrating the risk (uncertainty) in planning processes, modeling and taking into account human behaviors and flows.**

Objective We aim to give the robot some socially compliant behaviors by anticipating the near future (trajectories of mobile obstacle in the robot's surroundings) and by integrating knowledge from psychology, sociology and urban planning. In this context, we will focus on the following 3 topics.

Risk-based planning. Unlike static or controlled environments [0] where global path planning approaches are suitable, dealing with highly dynamic and uncertain environments requires to integrate the notion of risk (risk of collision, risk of disturbance). Then, we examine how motion planning approaches can integrate this risk in the generation and selection of the paths. An algorithm called RiskRRT was proposed in the previous eMotion team. This algorithm plans goal oriented trajectories that minimize the risk estimated at each instant. It fits environments that are highly dynamic and adapts to a representation of uncertainty [93] (see Figure 3 .a for illustration). Now, we extend this principle to be adapted to various risk evaluation methods (proposed in 3.2) and various situation (highways, urban environments, even in dense traffic).

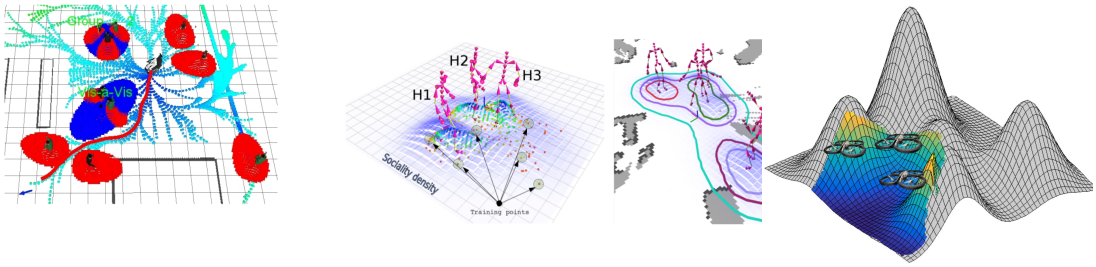


Figure 3. Illustrations of a. the Risk-RRT planning b. The human interaction space model c. Multi-UAV 3D coverage and exploration.

Recently we investigated the automatic learning of robot navigation in complex environments based on specific tasks and from visual input. We address this problem by combining computer vision, machine learning (deep-learning), and robotics path planning (see 7.5.1).

[0] Steven M. LaValle, Planning Algorithms, Cambridge University Press, 2006.
[0] known environment without uncertainty

Sharing the physical space with humans. Robots are expected to share their physical space with humans. Hence, robots need to take into account the presence of humans and to behave in a socially acceptable way. Their trajectories must be safe but also predictable, that is why they must follow social conventions, respecting proximity constraints, avoiding people interacting or joining a group engaged in conversation without disturbing. For this purpose, we proposed earlier to integrate some knowledge from the psychology domain (i.e. proxemics theory), see figure 3 .b. We aim now to integrate semantic knowledge [0] and psycho-social theories of human behavior [00] in the navigation framework we have developed for a few years (i.e. the Risk-based navigation algorithms [71], [93], [100]). These concepts were tested on our automated wheelchair (see figure 3 .c) but they have and will be adapted to autonomous cars, telepresence robots and companion robots. This work is currently supported by the ANR Valet and the ANR Hianic.

3.3.2. Decision Making in Multi-robot systems

Context. A central challenge in Chroma is to define **decision-making algorithms that scale up to large multi-robot systems**. This work takes place in the general framework of Multi-Agent Systems (MAS). The objective is to compute/define agent behaviors that provide cooperation and adaptation abilities. Solutions must also take into account the agent/robot computational limits.

We can abstract the challenge in three objectives :
i) mastering the complexity of large fleet of robots/vehicles (scalability),
ii) dealing with limited computational/memory capacity,
iii) building adaptive solutions (robustness).

Combining Decision-theoretic models and Swarm intelligence.

Over the past few years, our attempts to address multi-robot decision-making are mainly due to Multi-Agent Sequential Decision Making (MA-SDM) and Swarm Intelligence (SI). MA-SDM builds upon well-known decision-theoretic models (e.g., Markov decision processes and games) and related algorithms, that come with strong theoretical guarantees. In contrast, the expressiveness of MA-SDM models has limited scalability in face of realistic multi-robot systems [0], resulting in computational overload. On their side, SI methods, which rely on local rules – generally bio-inspired – and relating to Self-Organized Systems [0], can scale up to multiple robots and provide robustness to disturbances, but with poor theoretical guarantees [0]. Swarm models can also answer to the need of designing tractable solutions [92], but they remain not geared to express complex realistic tasks or to handle (point-to-point) communication between robots. This motivates our work to go beyond these two approaches and to combine them.

First, we plan to investigate **incremental expansion mechanisms in anytime decision-theoretic planning**, starting from local rules (from SI) to complex strategies with performance guarantees (from MA-SDM) [67]. This methodology is grounded into our research on anytime algorithms, that are guaranteed to stop at anytime while still providing a reliable solution to the original problem. It further relies on decision theoretical models and tools including: Decentralized and Partially Observable Markov Decision Processes and Games, Dynamic Programming, Distributed Reinforcement Learning and Statistical Machine Learning.

[0] B. Kuipers, The Spatial Semantic Hierarchy, Artificial Intelligence, Volume 119, Issues 1–2, May 2000, Pages 191-233
[0] Gibson, J. (1977). The theory of affordances, in Perceiving, Acting, and Knowing. Towards an Ecological Psychology. Number eds Shaw R., Bransford J. Hoboken,NJ: John Wiley & Sons Inc.
[0] Hall, E. (1966). The hidden dimension. Doubleday Anchor Books.
[0] Martin L. Puterman, Markov Decision Processes; Stuart Russell and Peter Norvig, Artificial Intelligence - A Modern Approach
[0] D. Floreano and C. Mattiussi, Bio-Inspired Artificial Intelligence - Theories, Methods, and Technologies, MIT Press, 2008.
[0] S. A. Brueckner, G. Di Marzo Serugendo, A. Karageorgos, R. Nagpal (2005). Engineering Self-Organising Systems, Methodologies and Applications. LNAI 3464 State-of-the-Art Survey, Springer book.

Second, we plan to extend the SI approach by considering **the integration of optimization techniques at the local level**. The purpose is to force the system to explore solutions around the current stabilized state – potentially a local optimum – of the system. We aim at keeping scalability and self-organization properties by not compromising the decentralized nature of such systems. Introducing optimization in this way requires to measure locally the performances, which is generally possible from local perception of robots (or using learning techniques). The main optimization methods we will consider are Local Search (Gradient Descent), Distributed Stochastic Algorithm and Reinforcement Learning. We have shown in [97] the interest of such an approach for driverless vehicle traffic optimization.

Both approaches must lead to **master the complexity** inherent to large and open multi-robot systems. Such systems are prone to combinatorial problems, in term of state space and communication, when the number of robots grows. To cope with this complexity we explore several approaches :

- Combining MA-SDM, machine learning and OR [0] techniques to deal with global-local optimization in multi-agent/robot systems. In 2016, we started a collaboration with the VOLVO Group, in Lyon, to deal with VRP problems and optimization of goods distribution using a fleet of autonomous vehicles. We also explore such a methodology in the framework of the collaboration with the team of Prof. G. Czibula (Cluj Univiersity, Romania).
- Defining heuristics by decentralizing global exact solutions. For instance we explore online stochastic-optimization planning to deal with multi-robot coverage/exploration of 3D environments, see Fig 3 .c and [42].

Beyond this methodological work, we aim to evaluate our models on benchmarks from the literature, by using simulation tools as a complement of robotic experiments. This will lead us to develop simulators, allowing to deploy tens to thousands robots in constrained environments.

Towards adaptive connected robots.

Mobile robots and autonomous vehicles are becoming more connected to one another and to other devices in the environment (concept of cloud of robots [0] and V2V/V2I connectivity in transportation systems). Such robotic systems are open systems as the number of connected entities is varying dynamically. Network of robots brought with them new problems, as the need of (online) adaption to changes in the system and to the variability of the communication.

In Chroma, we address the problem of adaptation by considering machine learning techniques and local mechanisms as discussed above (SI models). More specifically we investigate the problem of maintaining the connectivity between robots which perform dynamic version of tasks such as patrolling, exploration or transportation, i.e. where the setting of the problem is continuously changing and growing (see [86]).

In Lyon, the CITI Laboratory conducts research in many aspects of telecommunication, from signal theory to distributed computation. In this context, Chroma develops cooperations with the Inria team Agora [86] (wireless communication protocols) and with Dynamid team [63] (middlleware and cloud aspects), that we wish to reinforce in the next years.

[0] Operations Research

[0] see for instance the first International Workshop on Cloud and Robotics, 2016.

DEFROST Project-Team

3. Research Program

3.1. Introduction

Our research crosses different disciplines: numerical mechanics, control design, robotics, optimisation methods and clinical applications. Our organisation aims at facilitating the team work and cross-fertilisation of research results in the group. We have three objectives (1, 2 and 3) that correspond to the main scientific challenges. In addition, we have two transverse objectives that are also highly challenging: the development of a high performance software support for the project (objective 4) and the validation tools and protocols for the models and methods (objective 5).

3.2. Objective 1: Accurate model of soft robot deformation computed in finite time

The objective is to find concrete numerical solutions to the challenge of modelling soft robots with strong real-time constraints. To solve continuum mechanics equations, we will start our research with real-time FEM or equivalent methods that were developed for soft-tissue simulation. We will extend the functionalities to account for the needs of a soft-robotic system:

- Coupling with other physical phenomenons that govern the activity of sensors and actuators (hydraulic, pneumatic, electro-active polymers, shape-memory alloys...).
- Fulfilling the new computational time constraints (harder than surgical simulation for training) and find better tradeoff between cost and precision of numerical solvers using reduced-order modelling techniques with error control.
- Exploring interactive and semi-automatic optimisation methods for design based on obtained solution for fast computation on soft robot models.

3.3. Objective 2: Model based control of soft robot behavior

The focus of this objective is on obtaining a generic methodology for soft robot feedback control. Several steps are needed to design a model based control from FEM approach:

- The fundamental question of the kinematic link between actuators, sensors, effectors and contacts using the most reduced mathematical space must be carefully addressed. We need to find efficient algorithms for real-time projection of non-linear FEM models in order to pose the control problem using the only relevant parameters of the motion control.
- Intuitive remote control is obtained when the user directly controls the effector motion. To add this functionality, we need to obtain real-time inverse models of the soft robots by optimisation. Several criteria will be combined in this optimisation: effector motion control, structural stiffness of the robot, reduce intensity of the contact with the environment...
- Investigating closed-loop approaches using sensor feedback: as sensors cannot monitor all points of the deformable structure, the information provided will only be partial. We will need additional algorithms based on the FEM model to obtain the best possible treatment of the information. The final objective of these models and algorithms is to have robust and efficient feedback control strategies for soft robots. One of the main challenge here is to ensure / prove stability in closed-loop.

3.4. Objective 3: Modeling the interaction with a complex environment

Even if the inherent mechanical compliance of soft robots makes them safer, more robust and particularly adapted to interaction with fragile environments, the contact forces need to be controlled by:

- Setting up real-time modelling and the control methods needed to pilot the forces that the robot imposes on its environment and to control the robot deformations imposed by its environment. Note that if an operative task requires to apply forces on the surrounding structures, the robot must be anchored to other structures or structurally rigidified.
- Providing mechanics models of the environment that include the uncertainties on the geometry and on the mechanical properties, and are capable of being readjusted in real-time.
- Using the visual feedback of the robot behavior to adapt dynamically the models. The observation provided in the image coupled with an inverse accurate model of the robot could transform the soft robot into sensor: as the robot deforms with the contact of the surroundings, we could retrieve some missing parameters of the environment by a smart monitoring of the robot deformations.

3.5. Objective 4: Soft Robotics Software

Expected research results of this project are numerical methods and algorithms that require high-performance computing and suitability with robotic applications. There is no existing software support for such development. We propose to develop our own software, in a suite split into three applications:

- The first one will facilitate the design of deformable robots by an easy passage from CAD software (for the design of the robot) to the FEM based simulation.
- The second one is an anticipative clinical simulator. The aim is to co-design the robotic assistance with the physicians, thanks to a realistic simulation of the procedure or the robotic assistance. This will facilitate the work of reflection on new clinical approaches prior any manufacturing.
- The third one is the control design software. It will provide the real-time solutions for soft robot control developed in the project.

3.6. Objective 5: Validation and application demonstrations

The implementation of experimental validation is a key challenge for the project. On one side, we need to validate the model and control algorithms using concrete test case example in order to improve the modelling and to demonstrate the concrete feasibility of our methods. On the other side, concrete applications will also feed the reflexions on the objectives of the scientific program.

We will build our own experimental soft robots for the validation of objectives 2 and 3 when there is no existing “turn-key” solution. Designing and making our own soft robots, even if only for validation, will help the setting-up of adequate models.

For the validation of objective 4, we will develop “anatomical soft robot”: soft robot with the shape of organs, equipped with sensors (to measure the contact forces) and actuators (to be able to stiffen the walls and recreate natural motion of soft-tissues). We will progressively increase the level of realism of this novel validation set-up to come closer to the anatomical properties.

FLOWERS Project-Team

3. Research Program

3.1. Research Program

Research in artificial intelligence, machine learning and pattern recognition has produced a tremendous amount of results and concepts in the last decades. A blooming number of learning paradigms - supervised, unsupervised, reinforcement, active, associative, symbolic, connectionist, situated, hybrid, distributed learning... - nourished the elaboration of highly sophisticated algorithms for tasks such as visual object recognition, speech recognition, robot walking, grasping or navigation, the prediction of stock prices, the evaluation of risk for insurances, adaptive data routing on the internet, etc... Yet, we are still very far from being able to build machines capable of adapting to the physical and social environment with the flexibility, robustness, and versatility of a one-year-old human child.

Indeed, one striking characteristic of human children is the nearly open-ended diversity of the skills they learn. They not only can improve existing skills, but also continuously learn new ones. If evolution certainly provided them with specific pre-wiring for certain activities such as feeding or visual object tracking, evidence shows that there are also numerous skills that they learn smoothly but could not be "anticipated" by biological evolution, for example learning to drive a tricycle, using an electronic piano toy or using a video game joystick. On the contrary, existing learning machines, and robots in particular, are typically only able to learn a single pre-specified task or a single kind of skill. Once this task is learnt, for example walking with two legs, learning is over. If one wants the robot to learn a second task, for example grasping objects in its visual field, then an engineer needs to re-program manually its learning structures: traditional approaches to task-specific machine/robot learning typically include engineer choices of the relevant sensorimotor channels, specific design of the reward function, choices about when learning begins and ends, and what learning algorithms and associated parameters shall be optimized.

As can be seen, this requires a lot of important choices from the engineer, and one could hardly use the term "autonomous" learning. On the contrary, human children do not learn following anything looking like that process, at least during their very first years. Babies develop and explore the world by themselves, focusing their interest on various activities driven both by internal motives and social guidance from adults who only have a folk understanding of their brains. Adults provide learning opportunities and scaffolding, but eventually young babies always decide for themselves what activity to practice or not. Specific tasks are rarely imposed to them. Yet, they steadily discover and learn how to use their body as well as its relationships with the physical and social environment. Also, the spectrum of skills that they learn continuously expands in an organized manner: they undergo a developmental trajectory in which simple skills are learnt first, and skills of progressively increasing complexity are subsequently learnt.

A link can be made to educational systems where research in several domains have tried to study how to provide a good learning experience to learners. This includes the experiences that allow better learning, and in which sequence they must be experienced. This problem is complementary to that of the learner that tries to learn efficiently, and the teacher here has to use as efficiently the limited time and motivational resources of the learner. Several results from psychology [59] and neuroscience [85] have argued that the human brain feels intrinsic pleasure in practicing activities of optimal difficulty or challenge. A teacher must exploit such activities to create positive psychological states of flow [73].

A grand challenge is thus to be able to build machines that possess this capability to discover, adapt and develop continuously new know-how and new knowledge in unknown and changing environments, like human children. In 1950, Turing wrote that the child's brain would show us the way to intelligence: "Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's" [154]. Maybe, in opposition to work in the field of Artificial Intelligence who has focused on mechanisms trying to match the capabilities of "intelligent" human adults such as chess playing or natural language

dialogue [91], it is time to take the advice of Turing seriously. This is what a new field, called developmental (or epigenetic) robotics, is trying to achieve [109] [158]. The approach of developmental robotics consists in importing and implementing concepts and mechanisms from developmental psychology [117], cognitive linguistics [72], and developmental cognitive neuroscience [96] where there has been a considerable amount of research and theories to understand and explain how children learn and develop. A number of general principles are underlying this research agenda: embodiment [63] [131], grounding [89], situatedness [49], self-organization [150] [132], enaction [156], and incremental learning [67].

Among the many issues and challenges of developmental robotics, two of them are of paramount importance: exploration mechanisms and mechanisms for abstracting and making sense of initially unknown sensorimotor channels. Indeed, the typical space of sensorimotor skills that can be encountered and learnt by a developmental robot, as those encountered by human infants, is immensely vast and inhomogeneous. With a sufficiently rich environment and multimodal set of sensors and effectors, the space of possible sensorimotor activities is simply too large to be explored exhaustively in any robot's life time: it is impossible to learn all possible skills and represent all conceivable sensory percepts. Moreover, some skills are very basic to learn, some other very complicated, and many of them require the mastery of others in order to be learnt. For example, learning to manipulate a piano toy requires first to know how to move one's hand to reach the piano and how to touch specific parts of the toy with the fingers. And knowing how to move the hand might require to know how to track it visually.

Exploring such a space of skills randomly is bound to fail or result at best on very inefficient learning [128]. Thus, exploration needs to be organized and guided. The approach of epigenetic robotics is to take inspiration from the mechanisms that allow human infants to be progressively guided, i.e. to develop. There are two broad classes of guiding mechanisms which control exploration:

1. **internal guiding mechanisms,** and in particular intrinsic motivation, responsible of spontaneous exploration and curiosity in humans, which is one of the central mechanisms investigated in FLOWERS, and technically amounts to achieve online active self-regulation of the growth of complexity in learning situations;
2. **social learning and guidance,** a learning mechanisms that exploits the knowledge of other agents in the environment and/or that is guided by those same agents. These mechanisms exist in many different forms like emotional reinforcement, stimulus enhancement, social motivation, guidance, feedback or imitation, some of which being also investigated in FLOWERS;

3.1.1. Internal guiding mechanisms

In infant development, one observes a progressive increase of the complexity of activities with an associated progressive increase of capabilities [117], children do not learn everything at one time: for example, they first learn to roll over, then to crawl and sit, and only when these skills are operational, they begin to learn how to stand. The perceptual system also gradually develops, increasing children perceptual capabilities other time while they engage in activities like throwing or manipulating objects. This make it possible to learn to identify objects in more and more complex situations and to learn more and more of their physical characteristics.

Development is therefore progressive and incremental, and this might be a crucial feature explaining the efficiency with which children explore and learn so fast. Taking inspiration from these observations, some roboticists and researchers in machine learning have argued that learning a given task could be made much easier for a robot if it followed a developmental sequence and "started simple" [53] [79]. However, in these experiments, the developmental sequence was crafted by hand: roboticists manually build simpler versions of a complex task and put the robot successively in versions of the task of increasing complexity. And when they wanted the robot to learn a new task, they had to design a novel reward function.

Thus, there is a need for mechanisms that allow the autonomous control and generation of the developmental trajectory. Psychologists have proposed that intrinsic motivations play a crucial role. Intrinsic motivations are mechanisms that push humans to explore activities or situations that have intermediate/optimal levels of novelty, cognitive dissonance, or challenge [59] [73] [75]. The role and structure of intrinsic motivation in humans have been made more precise thanks to recent discoveries in neuroscience showing the implication

of dopaminergic circuits and in exploration behaviours and curiosity [74] [93] [147]. Based on this, a number of researchers have began in the past few years to build computational implementation of intrinsic motivation [128] [129] [145] [57] [94] [112] [146]. While initial models were developed for simple simulated worlds, a current challenge is to manage to build intrinsic motivation systems that can efficiently drive exploratory behaviour in high-dimensional unprepared real world robotic sensorimotor spaces [129], [128], [130], [143]. Specific and complex problems are posed by real sensorimotor spaces, in particular due to the fact that they are both high-dimensional as well as (usually) deeply inhomogeneous. As an example for the latter issue, some regions of real sensorimotor spaces are often unlearnable due to inherent stochasticity or difficulty, in which case heuristics based on the incentive to explore zones of maximal unpredictability or uncertainty, which are often used in the field of active learning [70] [90] typically lead to catastrophic results. The issue of high dimensionality does not only concern motor spaces, but also sensory spaces, leading to the problem of correctly identifying, among typically thousands of quantities, those latent variables that have links to behavioral choices. In FLOWERS, we aim at developing intrinsically motivated exploration mechanisms that scale in those spaces, by studying suitable abstraction processes in conjunction with exploration strategies.

3.1.2. Socially Guided and Interactive Learning

Social guidance is as important as intrinsic motivation in the cognitive development of human babies [117]. There is a vast literature on learning by demonstration in robots where the actions of humans in the environment are recognized and transferred to robots [52]. Most such approaches are completely passive: the human executes actions and the robot learns from the acquired data. Recently, the notion of interactive learning has been introduced in [151], [60], motivated by the various mechanisms that allow humans to socially guide a robot [139]. In an interactive context the steps of self-exploration and social guidance are not separated and a robot learns by self exploration and by receiving extra feedback from the social context [151], [99], [113].

Social guidance is also particularly important for learning to segment and categorize the perceptual space. Indeed, parents interact a lot with infants, for example teaching them to recognize and name objects or characteristics of these objects. Their role is particularly important in directing the infant attention towards objects of interest that will make it possible to simplify at first the perceptual space by pointing out a segment of the environment that can be isolated, named and acted upon. These interactions will then be complemented by the children own experiments on the objects chosen according to intrinsic motivation in order to improve the knowledge of the object, its physical properties and the actions that could be performed with it.

In FLOWERS, we are aiming at including intrinsic motivation system in the self-exploration part thus combining efficient self-learning with social guidance [122], [123]. We also work on developing perceptual capabilities by gradually segmenting the perceptual space and identifying objects and their characteristics through interaction with the user [110] and robots experiments [95]. Another challenge is to allow for more flexible interaction protocols with the user in terms of what type of feedback is provided and how it is provided [107].

Exploration mechanisms are combined with research in the following directions:

3.1.3. Cumulative learning, reinforcement learning and optimization of autonomous skill learning

FLOWERS develops machine learning algorithms that can allow embodied machines to acquire cumulatively sensorimotor skills. In particular, we develop optimization and reinforcement learning systems which allow robots to discover and learn dictionaries of motor primitives, and then combine them to form higher-level sensorimotor skills.

3.1.4. Autonomous perceptual and representation learning

In order to harness the complexity of perceptual and motor spaces, as well as to pave the way to higher-level cognitive skills, developmental learning requires abstraction mechanisms that can infer structural information out of sets of sensorimotor channels whose semantics is unknown, discovering for example the topology of the body or the sensorimotor contingencies (proprioceptive, visual and acoustic). This process is meant to

be open- ended, progressing in continuous operation from initially simple representations towards abstract concepts and categories similar to those used by humans. Our work focuses on the study of various techniques for:

- autonomous multimodal dimensionality reduction and concept discovery;
- incremental discovery and learning of objects using vision and active exploration, as well as of auditory speech invariants;
- learning of dictionaries of motion primitives with combinatorial structures, in combination with linguistic description;
- active learning of visual descriptors useful for action (e.g. grasping);

3.1.5. Embodiment and maturational constraints

FLOWERS studies how adequate morphologies and materials (i.e. morphological computation), associated to relevant dynamical motor primitives, can importantly simplify the acquisition of apparently very complex skills such as full-body dynamic walking in biped. FLOWERS also studies maturational constraints, which are mechanisms that allow for the progressive and controlled release of new degrees of freedoms in the sensorimotor space of robots.

3.1.6. Discovering and abstracting the structure of sets of uninterpreted sensors and motors

FLOWERS studies mechanisms that allow a robot to infer structural information out of sets of sensorimotor channels whose semantics is unknown, for example the topology of the body and the sensorimotor contingencies (proprioceptive, visual and acoustic). This process is meant to be open-ended, progressing in continuous operation from initially simple representations to abstract concepts and categories similar to those used by humans.

HEPHAISTOS Project-Team

3. Research Program

3.1. Interval analysis

We are interested in real-valued system solving ($f(X) = 0$, $f(X) \leq 0$), in optimization problems, and in the proof of the existence of properties (for example, it exists X such that $f(X) = 0$ or it exist two values X_1, X_2 such that $f(X_1) > 0$ and $f(X_2) < 0$). There are few restrictions on the function f as we are able to manage explicit functions using classical mathematical operators (e.g. $\sin(x + y) + \log(\cos(e^x) + y^2)$ as well as implicit functions (e.g. determining if there are parameter values of a parametrized matrix such that the determinant of the matrix is negative, without calculating the analytical form of the determinant).

Solutions are searched within a finite domain (called a *box*) which may be either continuous or mixed (i.e. for which some variables must belong to a continuous range while other variables may only have values within a discrete set). An important point is that we aim at finding all the solutions within the domain whenever the computer arithmetic will allow it: in other words we are looking for *certified* solutions. For example, for 0-dimensional system solving, we will provide a box that contains one, and only one, solution together with a numerical approximation of this solution. This solution may further be refined at will using multi-precision.

The core of our methods is the use of *interval analysis* that allows one to manipulate mathematical expressions whose unknowns have interval values. A basic component of interval analysis is the *interval evaluation* of an expression. Given an analytical expression F in the unknowns $\{x_1, x_2, ..., x_n\}$ and ranges $\{X_1, X_2, ..., X_n\}$ for these unknowns we are able to compute a range $[A, B]$, called the interval evaluation, such that

$$\forall \{x_1, x_2, ..., x_n\} \in \{X_1, X_2, ..., X_n\}, A \leq F(x_1, x_2, ..., x_n) \leq B \tag{1}$$

In other words the interval evaluation provides a lower bound of the minimum of F and an upper bound of its maximum over the box.

For example if $F = x\ sin(x + x^2)$ and $x \in [0.5, 1.6]$, then $F([0.5, 1.6]) = [-1.362037441, 1.6]$, meaning that for any x in [0.5,1.6] we guarantee that $-1.362037441 \leq f(x) \leq 1.6$.

The interval evaluation of an expression has interesting properties:

- it can be implemented in such a way that the results are guaranteed with respect to round-off errors i.e. property 1 is still valid in spite of numerical errors induced by the use of floating point numbers
- if $A > 0$ or $B < 0$, then no values of the unknowns in their respective ranges can cancel F
- if $A > 0$ ($B < 0$), then F is positive (negative) for any value of the unknowns in their respective ranges

A major drawback of the interval evaluation is that $A(B)$ may be overestimated i.e. values of $x_1, x_2, ..., x_n$ such that $F(x_1, x_2, ..., x_n) = A(B)$ may not exist. This overestimation occurs because in our calculation each occurrence of a variable is considered as an independent variable. Hence if a variable has multiple occurrences, then an overestimation may occur. Such phenomena can be observed in the previous example where $B = 1.6$ while the real maximum of F is approximately 0.9144. The value of B is obtained because we are using in our calculation the formula $F = xsin(y + z^2)$ with y, z having the same interval value as x.

Fortunately there are methods that allow one to reduce the overestimation and the overestimation amount decreases with the width of the ranges. The latter remark leads to the use of a branch-and-bound strategy in which for a given box a variable range will be bisected, thereby creating two new boxes that are stored in a list and processed later on. The algorithm is complete if all boxes in the list have been processed, or if during the process a box generates an answer to the problem at hand (e.g. if we want to prove that $F(X) < 0$, then the algorithm stops as soon as $F(\mathcal{B}) \geq 0$ for a certain box $\mathcal{B}$).

A generic interval analysis algorithm involves the following steps on the current box [8], [4]:

1. *exclusion operators*: these operators determine that there is no solution to the problem within a given box. An important issue here is the extensive and smart use of the monotonicity of the functions
2. *filters*: these operators may reduce the size of the box i.e. decrease the width of the allowed ranges for the variables
3. *existence operators*: they allow one to determine the existence of a unique solution within a given box and are usually associated with a numerical scheme that allows for the computation of this solution in a safe way
4. *bisection*: choose one of the variable and bisect its range for creating two new boxes
5. *storage*: store the new boxes in the list

The scope of the HEPHAISTOS project is to address all these steps in order to find the most efficient procedures. Our efforts focus on mathematical developments (adapting classical theorems to interval analysis, proving interval analysis theorems), the use of symbolic computation and formal proofs (a symbolic pre-processing allows one to automatically adapt the solver to the structure of the problem), software implementation and experimental tests (for validation purposes).

Important note: We have insisted on interval analysis because this is a **major component** or our robotics activity. Our theoretical work in robotics is an analysis of the robotic environment in order to exhibit proofs on the behavior of the system that may be qualitative (e.g. the proof that a cable-driven parallel robot with more than 6 non-deformable cables will have at most 6 cables under tension simultaneously) or quantitative. In the quantitative case as we are dealing with realistic and not toy examples (including our own prototypes that are developed whenever no equivalent hardware is available or to very our assumptions) we have to manage problems that are so complex that analytical solutions are probably out of reach (e.g. the direct kinematics of parallel robots) and we have to resort to algorithms and numerical analysis. We are aware of different approaches in numerical analysis (e.g. some team members were previously involved in teams devoted to computational geometry and algebraic geometry) but interval analysis provides us another approach with high flexibility, the possibility of managing non algebraic problems (e.g. the kinematics of cable-driven parallel robots with sagging cables, that involves inverse hyperbolic functions) and to address various types of issues (system solving, optimization, proof of existence ...). However whenever needed we will rely as well on continuation, algebraic geometry, geometry or learning.

3.2. Robotics

HEPHAISTOS, as a follow-up of COPRIN, has a long-standing tradition of robotics studies, especially for closed-loop robots [3], especially cable-driven parallel robots. We address theoretical issues with the purpose of obtaining analytical and theoretical solutions, but in many cases only numerical solutions can be obtained due to the complexity of the problem. This approach has motivated the use of interval analysis for two reasons:

1. the versatility of interval analysis allows us to address issues (e.g. singularity analysis) that cannot be tackled by any other method due to the size of the problem
2. uncertainties (which are inherent to a robotic device) have to be taken into account so that the *real* robot is guaranteed to have the same properties as the *theoretical* one, even in the worst case. This is a crucial issue for many applications in robotics (e.g. medical or assistance robot)

Our field of study in robotics focuses on *kinematic* issues such as workspace and singularity analysis, positioning accuracy, trajectory planning, reliability, calibration, modularity management and, prominently, *appropriate design*, i.e. determining the dimensioning of a robot mechanical architecture that guarantees that the real robot satisfies a given set of requirements. The methods that we develop can be used for other robotic problems, see for example the management of uncertainties in aircraft design [6].

Our theoretical work must be validated through experiments that are essential for the sake of credibility. A contrario, experiments will feed theoretical work. Hence HEPHAISTOS works with partners on the development of real robots but also develops its own prototypes. In the last years we have developed a large number of prototypes and we have extended our development to devices that are not strictly robots but are part of an overall environment for assistance. We benefit here from the development of new miniature, low energy computers with an interface for analog and logical sensors such as the Arduino or the Phidgets. The web pages http://www-sop.inria.fr/hephaistos/mediatheque/index.html presents all of our prototypes and experimental work.

LARSEN Project-Team

3. Research Program

3.1. Lifelong Autonomy

3.1.1. Scientific Context

So far, only a few autonomous robots have been deployed for a long time (weeks, months, or years) outside of factories and laboratories. They are mostly mobile robots that simply "move around" (e.g., vacuum cleaners or museum "guides") and data collecting robots (e.g., boats or underwater "gliders" that collect data about the water of the ocean).

A large part of the long-term autonomy community is focused on simultaneous localization and mapping (SLAM), with a recent emphasis on changing and outdoor environments [25], [34]. A more recent theme is life-long learning: during long-term deployment, we cannot hope to equip robots with everything they need to know, therefore some things will have to be learned along the way. Most of the work on this topic leverages machine learning and/or evolutionary algorithms to improve the ability of robots to react to unforeseen changes [25], [32].

3.1.2. Main Challenges

The first major challenge is to endow robots with a stable situation awareness in open and dynamic environments. This covers both the state estimation of the robot itself as well as the perception/representation of the environment. Both problems have been claimed to be solved but it is only the case for static environments [30].

In the LARSEN team, we aim at deployment in environments shared with humans which imply dynamic objects that degrade both the mapping and localization of a robot, especially in cluttered spaces. Moreover, when robots stay longer in the environment than for the acquisition of a snapshot map, they have to face structural changes, such as the displacement of a piece of furniture or the opening or closing of a door. The current approach is to simply update an implicitly static map with all observations with no attempt at distinguishing the suitable changes. For localization in not-too-cluttered or not-too-empty environments, this is generally sufficient as a significant fraction of the environment should remain stable. But for life-long autonomy, and in particular navigation, the quality of the map, and especially the knowledge of the stable parts, is primordial.

A second major obstacle to move robots outside of labs and factories is their fragility: Current robots often break in a few hours, if not a few minutes. This fragility mainly stems from the overall complexity of robotic systems, which involve many actuators, many sensors, and complex decisions, and from the diversity of situations that robots can encounter. Low-cost robots exacerbate this issue because they can be broken in many ways (high-quality material is expensive), because they have low self-sensing abilities (sensors are expensive and increase the overall complexity), and because they are typically targeted towards non-controlled environments (e.g., houses rather than factories, in which robots are protected from most unexpected events). More generally, this fragility is a symptom of the lack of adaptive abilities in current robots.

3.1.3. Angle of Attack

To solve the state estimation problem, our approach is to combine classical estimation filters (Extended Kalman Filters, Unscented Kalman Filters, or particle filters) with a Bayesian reasoning model in order to internally simulate various configurations of the robot in its environment. This should allow for adaptive estimation that can be used as one aspect of long-term adaptation. To handle dynamic and structural changes in an environment, we aim at assessing, for each piece of observation, whether it is static or not.

We also plan to address active sensing to improve the situation awareness of robots. Literally, active sensing is the ability of an interacting agent to act so as to control what it senses from its environment with the typical objective of acquiring information about this environment. A formalism for representing and solving active sensing problems has already been proposed by members of the team [24] and we aim to use this to formalize decision making problems of improving situation awareness.

Situation awareness of robots can also be tackled by cooperation, whether it be between robots or between robots and sensors in the environment (led out intelligent spaces) or between robots and humans. This is in rupture with classical robotics, in which robots are conceived as self-contained. But, in order to cope with as diverse environments as possible, these classical robots use precise, expensive, and specialized sensors, whose cost prohibits their use in large-scale deployments for service or assistance applications. Furthermore, when all sensors are on the robot, they share the same point of view on the environment, which is a limit for perception. Therefore, we propose to complement a cheaper robot with sensors distributed in a target environment. This is an emerging research direction that shares some of the problematics of multi-robot operation and we are therefore collaborating with other teams at Inria that address the issue of communication and interoperability.

To address the fragility problem, the traditional approach is to first diagnose the situation, then use a planning algorithm to create/select a contingency plan. But, again, this calls for both expensive sensors on the robot for the diagnosis and extensive work to predict and plan for all the possible faults that, in an open and dynamic environment, are almost infinite. An alternative approach is then to skip the diagnosis and let the robot discover by trial and error a behavior that works in spite of the damage with a reinforcement learning algorithm [39], [32]. However, current reinforcement learning algorithms require hundreds of trials/episodes to learn a single, often simplified, task [32], which makes them impossible to use for real robots and more ambitious tasks. We therefore need to design new trial-and-error algorithms that will allow robots to learn with a much smaller number of trials (typically, a dozen). We think the key idea is to guide online learning on the physical robot with dynamic simulations. For instance, in our recent work, we successfully mixed evolutionary search in simulation, physical tests on the robot, and machine learning to allow a robot to recover from physical damage [33], [1].

A final approach to address fragility is to deploy several robots or a swarm of robots or to make robots evolve in an active environment. We will consider several paradigms such as (1) those inspired from collective natural phenomena in which the environment plays an active role for coordinating the activity of a huge number of biological entities such as ants and (2) those based on online learning [29]. We envision to transfer our knowledge of such phenomenon to engineer new artificial devices such as an intelligent floor (which is in fact a spatially distributed network in which each node can sense, compute and communicate with contiguous nodes and can interact with moving entities on top of it) in order to assist people and robots (see the principle in [37], [29], [23]).

3.2. Natural Interaction with Robotic Systems

3.2.1. Scientific Context

Interaction with the environment is a primordial requirement for an autonomous robot. When the environment is sensorized, the interaction can include localizing, tracking, and recognizing the behavior of robots and humans. One specific issue lies in the lack of predictive models for human behavior and a critical constraint arises from the incomplete knowledge of the environment and the other agents.

On the other hand, when working in the proximity of or directly with humans, robots must be capable of safely interacting with them, which calls upon a mixture of physical and social skills. Currently, robot operators are usually trained and specialized but potential end-users of robots for service or personal assistance are not skilled robotics experts, which means that the robot needs to be accepted as reliable, trustworthy and efficient [42]. Most Human-Robot Interaction (HRI) studies focus on verbal communication [38] but applications such as assistance robotics require a deeper knowledge of the intertwined exchange of social and physical signals to provide suitable robot controllers.

3.2.2. Main Challenges

We are here interested in building the bricks for a situated Human-Robot Interaction (HRI) addressing both the physical and social dimension of the close interaction, and the cognitive aspects related to the analysis and interpretation of human movement and activity.

The combination of physical and social signals into robot control is a crucial investigation for assistance robots [40] and robotic co-workers [36]. A major obstacle is the control of physical interaction (precisely, the control of contact forces) between the robot and the human while both partners are moving. In mobile robots, this problem is usually addressed by planning the robot movement taking into account the human as an obstacle or as a target, then delegating the execution of this "high-level" motion to whole-body controllers, where a mixture of weighted tasks is used to account for the robot balance, constraints, and desired end-effector trajectories [26].

The first challenge is to make these controllers easier to deploy in real robotics systems, as currently they require a lot of tuning and can become very complex to handle the interaction with unknown dynamical systems such as humans. Here, the key is to combine machine learning techniques with such controllers.

The second challenge is to make the robot react and adapt online to the human feedback, exploiting the whole set of measurable verbal and non-verbal signals that humans naturally produce during a physical or social interaction. Technically, this means finding the optimal policy that adapts the robot controllers online, taking into account feedback from the human. Here, we need to carefully identify the significant feedback signals or some metrics of human feedback. In real-world conditions (i.e., outside the research laboratory environment) the set of signals is technologically limited by the robot's and environmental sensors and the onboard processing capabilities.

The third challenge is for a robot to be able to identify and track people on board. The motivation is to be able to estimate online either the position, the posture, or even moods and intentions of persons surrounding the robot. The main challenge is to be able to do that online, in real-time and in cluttered environments.

3.2.3. Angle of Attack

Our key idea is to exploit the physical and social signals produced by the human during the interaction with the robot and the environment in controlled conditions, to learn simple models of human behavior and consequently to use these models to optimize the robot movements and actions. In a first phase, we will exploit human physical signals (e.g., posture and force measurements) to identify the elementary posture tasks during balance and physical interaction. The identified model will be used to optimize the robot whole-body control as prior knowledge to improve both the robot balance and the control of the interaction forces. Technically, we will combine weighted and prioritized controllers with stochastic optimization techniques. To adapt online the control of physical interaction and make it possible with human partners that are not robotics experts, we will exploit verbal and non-verbal signals (e.g., gaze, touch, prosody). The idea here is to estimate online from these signals the human intent along with some inter-individual factors that the robot can exploit to adapt its behavior, maximizing the engagement and acceptability during the interaction.

Another promising approach already investigated in the LARSEN team is the capability for a robot and/or an intelligent space to localize humans in its surrounding environment and to understand their activities. This is an important issue to handle both for safe and efficient human-robot interaction.

Simultaneous Tracking and Activity Recognition (STAR) [41] is an approach we want to develop. The activity of a person is highly correlated with his position, and this approach aims at combining tracking and activity recognition to benefit one from another. By tracking the individual, the system may help infer its possible activity, while by estimating the activity of the individual, the system may make a better prediction of his/her possible future positions (especially in the case of occlusions). This direction has been tested with simulator and particle filters [28], and one promising direction would be to couple STAR with decision making formalisms like partially observable Markov decision processes (POMDPs). This would allow us to formalize problems such as deciding which action to take given an estimate of the human location and activity. This could also formalize other problems linked to the active sensing direction of the team: how the robotic system

should choose its actions in order to have a better estimate of the human location and activity (for instance by moving in the environment or by changing the orientation of its cameras)?

Another issue we want to address is robotic human body pose estimation. Human body pose estimation consists of tracking body parts by analyzing a sequence of input images from single or multiple cameras.

Human posture analysis is of high value for human robot interaction and activity recognition. However, even if the arrival of new sensors like RGB-D cameras has simplified the problem, it still poses a great challenge, especially if we want to do it online, on a robot and in realistic world conditions (cluttered environment). This is even more difficult for a robot to bring together different capabilities both at the perception and navigation level [27]. This will be tackled through different techniques, going from Bayesian state estimation (particle filtering), to learning, active and distributed sensing.

PERVASIVE Project-Team

3. Research Program

3.1. Modelling Human Awareness and Undestanding

The objectives of this research area are to develop and refine new computational techniques that improve the reliability and performance of situation models, extend the range of possible application domains, and reduce the cost of developing and maintaining situation models. Important research challenges include developing machine-learning techniques to automatically acquire and adapt situation models through interaction, development of techniques to reason and learn about appropriate behaviors, and the development of new algorithms and data structures for representing situation models.

Pervasive has addressed the following research challenges:

Techniques for learning and adapting situation models: Hand crafting of situation models is currently an expensive process requiring extensive trial and error. We will investigate combination of interactive design tools coupled with supervised and semi-supervised learning techniques for constructing initial, simplified prototype situation models in the laboratory. One possible approach is to explore developmental learning to enrich and adapt the range of situations and behaviors through interaction with users.

Reasoning about actions and behaviors: Constructing systems for reasoning about actions and their consequences is an important open challenge. We will explore integration of planning techniques for operationalizing actions sequences within behaviors, and for constructing new action sequences when faced with unexpected difficulties. We will also investigate reasoning techniques within the situation modeling process for anticipating the consequences of actions, events and phenomena.

Algorithms and data structures for situation models: In recent years, we have experimented with an architecture for situated interaction inspired by work in human factors. This model organises perception and interaction as a cyclic process in which directed perception is used to detect and track entities, verify relations between entities, detect trends, anticipate consequences and plan actions. Each phase of this process raises interesting challenges questions algorithms and programming techniques. We will experiment alternative programming techniques representing and reasoning about situation models both in terms of difficulty of specification and development and in terms of efficiency of the resulting implementation. We will also investigate the use of probabilistic graph models as a means to better accommodate uncertain and unreliable information. In particular, we will experiment with using probabilistic predicates for defining situations, and maintaining likelihood scores over multiple situations within a context. Finally, we will investigate the use of simulation as technique for reasoning about consequences of actions and phenomena.

The challenges in this research area have been addressed through three specific research actions covering situation modelling in homes, learning on mobile devices, and reasoning in critical situations.

3.1.1. Learning Routine patterns of activity in the home.

The objective of this research action is to develop a scalable approach to learning routine patterns of activity in a home using situation models. Information about user actions is used to construct situation models in which key elements are semantic representations of time, place, social role and actions. Activities are encoded as sequences of situations. Recurrent activities are detected as sequences of activities that occur at a specific time and place each day. Recurrent activities provide routines what can be used to predict future actions and anticipate needs and services. An early demonstration has been to construct an intelligent assistant that can respond to and filter communications.

This research action is carried out as part of the doctoral research of Julien Cumin in cooperation with researchers at Orange labs, Meylan. Results are to be published at Ubicomp, Ambient intelligence, Intelligent Environments and IEEE Transactions on System Man and Cybernetics. Julien Cumin will complete and defend his doctoral thesis in 2018.

3.1.2. Learning Patterns of Activity with Mobile Devices

The objective of this research action is to develop techniques to observe and learn recurrent patterns of activity using the full suite of sensors available on mobile devices such as tablets and smart phones. Most mobile devices include seven or more sensors organized in 4 groups: Positioning Sensors, Environmental Sensors, Communications Subsystems, and Sensors for Human-Computer Interaction. Taken together, these sensors can provide a very rich source of information about individual activity.

In this area we explore techniques to observe activity with mobiles devices in order to learn daily patterns of activity. We will explore supervised and semi-supervised learning to construct systems to recognize places and relevant activities. Location and place information, semantic time of day, communication activities, interpersonal interactions, and travel activities (walking, driving, riding public transportation, etc.) are recognized as probabilistic predicates and used to construct situation models. Recurrent sequences of situations will be detected and recorded to provide an ability to predict upcoming situations and anticipate needs for information and services.

Our goal is to develop a theory for building context aware services that can be deployed as part of the mobile applications that companies such as SNCF and RATP use to interact with clients. For example, a current project concerns systems that observe daily travel routines for the Paris region RATP metro and SNCF commuter trains. This system learns individual travel routines on the mobile device without the need to divulge information about personal travel to a cloud based system. The resulting service will consult train and metro schedules to assure that planned travel is feasible and to suggest alternatives in the case of travel disruptions. Similar applications are under discussion for the SNCF inter-city travel and Air France for air travel.

This research action is conducted in collaboration with the Inria Startup Situ8ed. The current objective is to deploy and evaluate a first prototype App during 2017. Techniques will be used commercially by Situ8ed for products to be deployed as early as 2019.

3.1.3. Bibliography

[Brdiczka 07] O. Brdiczka, "Learning Situation Models for Context-Aware Services", Doctoral Thesis of the INPG, 25 may 2007.

[Barraquand 12] R. Barraquand, "Design of Sociable Technologies", Doctoral Thesis of the University Grenoble Alps, 2 Feb 2012.

3.2. Perception of People, Activities and Emotions

Machine perception is fundamental for situated behavior. Work in this area concerns construction of perceptual components using computer vision, acoustic perception, accelerometers and other embedded sensors. These include low-cost accelerometers [Bao 04], gyroscopic sensors and magnetometers, vibration sensors, electromagnetic spectrum and signal strength (wifi, bluetooth, GSM), infrared presence detectors, and bolometric imagers, as well as microphones and cameras. With electrical usage monitoring, every power switch can be used as a sensor [Fogarty 06], [Coutaz 16]. We have developed perceptual components for integrated vision systems that combine a low-cost imaging sensors with on-board image processing and wireless communications in a small, low-cost package. Such devices are increasingly available, with the enabling manufacturing technologies driven by the market for integrated imaging sensors on mobile devices. Such technology enables the use of embedded computer vision as a practical sensor for smart objects.

Research challenges addressed in this area include development of practical techniques that can be deployed on smart objects for perception of people and their activities in real world environments, integration and fusion of information from a variety of sensor modalities with different response times and levels of abstraction, and perception of human attention, engagement, and emotion using visual and acoustic sensors.

Work in this research area will focus on three specific Research Actions

3.2.1. Multi-modal perception and modelling of activities

The objective of this research action is to develop techniques for observing and scripting activities for common household tasks such as cooking and cleaning. An important part of this project involves acquiring annotated multi-modal datasets of activity using an extensive suite of visual, acoustic and other sensors. We are interested in real-time on-line techniques that capture and model full body movements, head motion and manipulation actions as 3D articulated motion sequences decorated with semantic labels for individual actions and activities with multiple RGB and RGB-D cameras.

We have explored the integration of 3D articulated models with appearance based recognition approaches and statistical learning for modeling behaviors. Such techniques provide an important enabling technology for context aware services in smart environments [Coutaz 05], [Crowley 15], investigated by Pervasive Interaction team, as well as research on automatic cinematography and film editing investigated by the Imagine team [Gandhi 13] [Gandhi 14] [Ronfard 14] [Galvane 15]. An important challenge is to determine which techniques are most appropriate for detecting, modeling and recognizing a large vocabulary of actions and activities under different observational conditions.

We explored representations of behavior that encodes both temporal-spatial structure and motion at multiple levels of abstraction. We will further propose parameters to encode temporal constraints between actions in the activity classification model using a combination of higher-level action grammars [Pirsiavash 14] and episodic reasoning [Santofimia 14] [Edwards 14].

We have adapted this work to construct narrative descriptions of cooking activities from ego-centric vision, in cooperation with Remi Ronfard of the Imagine Team of Inria.

3.2.2. Perception with low-cost integrated sensors

In this research action, we will continue work on low-cost integrated sensors using visible light, infrared, and acoustic perception. We will continue development of integrated visual sensors that combine micro-cameras and embedded image processing for detecting and recognizing objects in storage areas. We will combine visual and acoustic sensors to monitor activity at work-surfaces. Low cost real-time image analysis procedures will be designed that acquire and process images directly as they are acquired by the sensor.

Bolometric image sensors measure the Far Infrared emissions of surfaces in order to provide an image in which each pixel is an estimate of surface temperature. Within the European MIRTIC project, Grenoble startup, ULIS has created a relatively low-cost Bolometric image sensor (Retina) that provides small images of 80 by 80 pixels taken from the Far-infrared spectrum. Each pixel provides an estimate of surface temperature. Working with Schneider Electric, engineers in the Pervasive Interaction team had developed a small, integrated sensor that combines the MIRTIC Bolometric imager with a microprocessor for on-board image processing. The package has been equipped with a fish-eye lens so that an overhead sensor mounted at a height of 3 meters has a field of view of approximately 5 by 5 meters. Real-time algorithms have been demonstrated for detecting, tracking and counting people, estimating their trajectories and work areas, and estimating posture.

Many of the applications scenarios for Bolometric sensors proposed by Schneider Electric assume a scene model that assigns pixels to surfaces of the floor, walls, windows, desks or other items of furniture. The high cost of providing such models for each installation of the sensor would prohibit most practical applications. We have recently developed a novel automatic calibration algorithm that determines the nature of the surface under each pixel of the sensor.

Work in this area will continue to develop low-cost real time infrared image sensing, as well as explore combinations of far-infrared images with RGB and RGBD images.

3.2.3. Observing and Modelling Competence and Awareness from Eye-gaze and Emotion

Humans display awareness and emotions through a variety of non-verbal channels. It is increasingly possible to record and interpret such information with available technology. Publicly available software can be used to efficiently detect and track face orientation using web cameras. Concentration can be inferred from changes in pupil size [Kahneman 66]. Observation of Facial Action Units [Ekman 71] can be used to detect both sustained

and instantaneous (micro-expressions) displays of valence and excitation. Heart rate can be measured from the Blood Volume Pulse as observed from facial skin color [Poh 11]. Body posture and gesture can be obtained from low-cost RGB sensors with depth information (RGB+D) [Shotton 13] or directly from images using detectors learned using deep learning [Ramakrishna 14]. Awareness and attention can be inferred from eye-gaze (scan path) and fixation using eye-tracking glasses as well as remote eye tracking devices [Holmqvist 11]. Such recordings can be used to reveal awareness of the current situation and to predict ability to respond effectively to opportunities and threats.

This work is supported by the ANR project CEEGE in cooperation with the department of NeuroCognition of Univ. Bielefeld. Work in this area includes the Doctoral research of Thomas Guntz to be defended in 2019.

3.2.4. Bibliography

[Bao 04] L. Bao, and S. S. Intille. "Activity recognition from user-annotated acceleration data.", IEEE Pervasive computing. Springer Berlin Heidelberg, pp. 1-17, 2004.

[Fogarty 06] J. Fogarty, C. Au and S. E. Hudson. "Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition." In Proceedings of the 19th annual ACM symposium on User interface software and technology, UIST 2006, pp. 91-100. ACM, 2006.

[Coutaz 16] J. Coutaz and J.L. Crowley, A First-Person Experience with End-User Development for Smart Homes. IEEE Pervasive Computing, 15(2), pp. 26-39, 2016.

[Coutaz 05] J. Coutaz, J.L. Crowley, S. Dobson, D. Garlan, "Context is key", Communications of the ACM, 48 (3), pp. 49-53, 2005.

[Crowley 15] J. L. Crowley and J. Coutaz, "An Ecological View of Smart Home Technologies", 2015 European Conference on Ambient Intelligence, Athens, Greece, Nov. 2015.

[Gandhi 13] Vineet Gandhi, Remi Ronfard. "Detecting and Naming Actors in Movies using Generative Appearance Models", Computer Vision and Pattern Recognition, 2013.

[Gandhi 14] Vineet Gandhi, Rémi Ronfard, Michael Gleicher. "Multi-Clip Video Editing from a Single Viewpoint", European Conference on Visual Media Production, 2014

[Ronfard 14] R. Ronfard, N. Szilas. "Where story and media meet: computer generation of narrative discourse". Computational Models of Narrative, 2014.

[Galvane 15] Quentin Galvane, Rémi Ronfard, Christophe Lino, Marc Christie. "Continuity Editing for 3D Animation". AAAI Conference on Artificial Intelligence, Jan 2015.

[Pirsiavash 14] Hamed Pirsiavash , Deva Ramanan, "Parsing Videos of Actions with Segmental Grammars", Computer Vision and Pattern Recognition, pp. 612-619, 2014.

[Edwards 14] C. Edwards. 2014, "Decoding the language of human movement". Commun. ACM 57, 12, pp. 12-14, November 2014.

[Kahneman 66] D. Kahneman and J. Beatty, Pupil diameter and load on memory. Science, 154(3756), pp. 1583-1585, 1966.

[Ekman 71] P. Ekman and W.V. Friesen, Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124, 1971.

[Poh 11] M. Z. Poh, D. J. McDuff, and R. W. Picard, Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans. Biomed. Eng., 58, pp. 7–11, 2011.

[Shotton 13] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, Real-time human pose recognition in parts from single depth images. Commun. ACM, 56, pp. 116–124, 2013.

[Ramakrishna 14] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell and Y. Sheikh, Pose machines: Articulated pose estimation via inference machines. In European Conference on Computer Vision (ECCV 2016), pp. 33-47, Springer, 2014.

[Cao 17] Z. Cao, T. Simon, S. E. Wei and Y. Sheikh, Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), IEEE Press, pp. 1302-1310, July, 2017.

[Holmqvist 11] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer, Eye Tracking: A Comprehensive Guide to Methods and Measures, OUP Oxford: Oxford, UK, 2011.

3.3. Sociable Interaction with Smart Objects

Reeves and Nass argue that a social interface may be the truly universal interface [Reeves 98]. Current systems lack ability for social interaction because they are unable to perceive and understand humans or to learn from interaction with humans. One of the goals of the research to be performed in Pervasive Interaction is to provide such abilities.

Work in research area RA3 will demonstrate the use of situation models for sociable interaction with smart objects and companion robots. We will explore the use of situation models as a representation for sociable interaction. Our goal in this research is to develop methods to endow an artificial agent with the ability to acquire social common sense using the implicit feedback obtained from interaction with people. We believe that such methods can provide a foundation for socially polite man-machine interaction, and ultimately for other forms of cognitive abilities. We propose to capture social common sense by training the appropriateness of behaviors in social situations. A key challenge is to employ an adequate representation for social situations.

Knowledge for sociable interaction will be encoded as a network of situations that capture both linguistic and non-verbal interaction cues and proper behavioral responses. Stereotypical social interactions will be represented as trajectories through the situation graph. We will explore methods that start from simple stereotypical situation models and extending a situation graph through the addition of new situations and the splitting of existing situations. An important aspect of social common sense is the ability to act appropriately in social situations. We propose to learn the association between behaviors and social situation using reinforcement learning. Situation models will be used as a structure for learning appropriateness of actions and behaviors that may be chosen in each situation, using reinforcement learning to determine a score for appropriateness based on feedback obtained by observing partners during interaction.

Work in this research area will focus on four specific Research Actions

3.3.1. Moving with people

Our objective in this area is to establish the foundations for robot motions that are aware of human social situation that move in a manner that complies with the social context, social expectations, social conventions and cognitive abilities of humans. Appropriate and socially compliant interactions require the ability for real time perception of the identity, social role, actions, activities and intents of humans. Such perception can be used to dynamically model the current situation in order to understand the situation and to compute the appropriate course of action for the robot depending on the task at hand.

To reach this objective, we propose to investigate three interacting research areas:

- Modeling the context and situation of human activities for motion planning
- Planning and acting in a social context.
- Identifying and modeling interaction behaviors.

In particular, we will investigate techniques that allow a tele-presence robot, such as the BEAM system, to autonomously navigate in crowds of people as may be found at the entry to a conference room, or in the hallway of a scientific meeting.

3.3.2. Understanding and communicating intentions from motion

This research area concerns the communication through motion. When two or more people move as a group, their motion is regulated by implicit rules that signal a shared sense of social conventions and social roles. For example, moving towards someone while looking directly at them signals an intention for engagement. In

certain cultures, subtle rules dictate who passes through a door first or last. When humans move in groups, they implicitly communicate intentions with motion. In this research area, we will explore the scientific literature on proxemics and the social sciences on such movements, in order to encode and evaluate techniques for socially appropriate motion by robots.

3.3.3. Socially aware interaction

This research area concerns socially aware man-machine interaction. Appropriate and socially compliant interaction requires the ability for real time perception of the identity, social role, actions, activities and intents of humans. Such perception can be used to dynamically model the current situation in order to understand the context and to compute the appropriate course of action for the task at hand. Performing such interactions in manner that respects and complies with human social norms and conventions requires models for social roles and norms of behavior as well as the ability to adapt to local social conventions and individual user preferences. In this research area, we will complement research area 3.2 with other forms of communication and interaction, including expression with stylistic face expressions rendered on a tablet, facial gestures, body motions and speech synthesis. We will experiment with use of commercially available tool for spoken language interaction in conjunction with expressive gestures.

3.3.4. Stimulating affection and persuasion with affective devices.

This research area concerns technologies that can stimulate affection and engagement, as well as induce changes in behavior. When acting as a coach or cooking advisor, smart objects must be credible and persuasive. One way to achieve this goal is to express affective feedbacks while interacting. This can be done using sound, light and/or complex moves when the system is composed of actuators.

Research in this area will address 3 questions:

1. How do human perceive affective signals expressed by smart objects (including robots)?
2. How does physical embodiment effect perception of affect by humans?
3. What are the most effective models and tools for animation of affective expression?

Both the physical form and the range of motion have important impact on the ability of a system to inspire affection. We will create new models to propose a generic animation model, and explore the effectiveness of different forms of motion in stimulating affect.

3.3.5. Bibliography

[Reeves 98] B. Reeves and C. Nass, The Media Equation: how People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press, 1998.

3.4. Interaction with Pervasive Smart Objects and Displays

Currently, the most effective technologies for new media for sensing, perception and experience are provided by virtual and augmented realities [Van Krevelen 2010]. At the same time, the most effective means to augment human cognitive abilities are provided by access to information spaces such as the world-wide-web using graphical user interfaces. A current challenge is to bring these two media together.

Display technologies continue to decrease exponentially, driven largely by investment in consumer electronics as well as the overall decrease in cost of microelectronics. A consequence has been an increasing deployment of digital displays in both public and private spaces. This trend is likely to accelerate, as new technologies and growth in available communications bandwidth enable ubiquitous low-cost access to information and communications.

The arrival of pervasive displays raises a number of interesting challenges for situated multi-modal interaction. For example:

1. Can we use perception to detect user engagement and identify users in public spaces?
2. Can we replace traditional pointing hardware with gaze and gesture based interaction?
3. Can we tailor information and interaction for truly situated interaction, providing the right information at the right time using the right interaction modality?
4. How can we avoid information overload and unnecessary distraction with pervasive displays?

It is increasingly possible to embed sensors and displays in clothing and ordinary devices, leading to new forms of tangible and wearable interaction with information. This raises challenges such as

1. What are the tradeoffs between large-scale environmental displays and wearable displays using technologies such as e-textiles and pico-projector?
2. How can we manage the tradeoffs between implicit and explicit interaction with both tangible and wearable interaction?
3. How can we determine the appropriate modalities for interaction?
4. How can we make users aware of interaction possibilities without creating distraction?

In addition to display and communications, the continued decrease in microelectronics has also driven an exponential decrease in cost of sensors, actuators, and computing resulting in an exponential growth in the number of smart objects in human environments. Current models for systems organization are based on centralized control, in which a controller or local hub, orchestrates smart objects, generally in connection with cloud computing. This model creates problems with privacy and ownership of information. An alternative is to organize local collections of smart objects to provide distributed services without the use of a centralized controller. The science of ecology can provide an architectural model for such organization.

This approach raises a number of interesting research challenges for pervasive interaction:

1. Can we devise distributed models for multi-modal fusion and interaction with information on heterogeneous devices?
2. Can we devise models for distributed interaction that migrates over available devices as the user changes location and task?
3. Can we manage migration of interaction over devices in a manner that provides seamless immersive interaction with information, services and media?
4. Can we provide models of distributed interaction that conserve the interaction context as services migrate?

Research Actions for Interaction with Pervasive Smart Objects for the period 2017 - 2020 include

3.4.1. Wearable and tangible interaction with smart textiles and wearable projectors

Opportunities in this area result from the emergence of new forms of interactive media using smart objects. We will explore the use of smart objects as tangible interfaces that make it possible to experience and interact with information and services by grasping and manipulating objects. We will explore the use of sensors and actuators in clothing and wearable devices such as gloves, hats and wrist bands both as a means of unobtrusively sensing human intentions and emotional states and as a means of stimulating human senses through vibration and sound. We will explore the new forms of interaction and immersion made possible by deploying interactive displays over large areas of an environment.

3.4.2. Pervasive interaction with ecologies of smart objects in the home

In this research area, we will explore and evaluate interaction with ecologies of smart objects in home environments. We will explore development of a range of smart objects that provide information services, such as devices for Episodic Memory for work surfaces and storage areas, devices to provide energy efficient control of environmental conditions, and interactive media that collect and display information. We propose to develop a new class of socially aware managers that coordinate smart objects and manage logistics in functional areas such as the kitchen, living rooms, closets, bedrooms, bathroom or office.

3.4.3. Bibliography

[Van Krevelen 10] D. W. F. Van Krevelen and R. Poelman, A survey of augmented reality technologies, applications and limitations. International Journal of Virtual Reality, 9(2), 1-20, 2010

RAINBOW Project-Team

3. Research Program

3.1. Main Vision

The vision of Rainbow (and foreseen applications) calls for several general scientific challenges: (i) high-level of autonomy for complex robots in complex (unstructured) environments, (ii) forward interfaces for letting an operator giving high-level commands to the robot, (iii) backward interfaces for informing the operator about the robot 'status', (iv) user studies for assessing the best interfacing, which will clearly depend on the particular task/situation. Within Rainbow we plan to tackle these challenges at different levels of depth:

- the **methodological and algorithmic side** of the sought human-robot interaction will be the **main focus** of Rainbow. Here, we will be interest in advancing the state-of-the-art in sensor-based online planning, control and manipulation for mobile/fixed robots. For instance, while classically most control approaches (especially those sensor-based) have been essentially *reactive*, we believe that less myopic strategies based on online/reactive trajectory optimization will be needed for the future Rainbow activities. The core ideas of Model-Predictive Control approaches (also known as Receding Horizon) or, in general, numerical optimal control methods will play a role in the Rainbow activities, for allowing the robots to reason/plan over some future time window and better cope with constraints. We will also consider extending classical sensor-based motion control/manipulation techniques to more realistic scenarios, such as deformable/flexible objects ("**Advanced Sensor-based Control**" axis). Finally, it will also be important to spend research efforts into the field of *Optimal Sensing*, in the sense of generating (again) trajectories that can optimize the state estimation problem in presence of scarce sensory inputs and/or non-negligible measurement and process noises, especially true for the case of mobile robots ("**Optimal and Uncertainty-Aware Sensing**" axis). We also aim at addressing the case of coordination between a single human user and multiple robots where, clearly, as explained the autonomy part plays even a more crucial role (no human can control multiple robots at once, thus a high degree of autonomy will be required by the robot group for executing the human commands);
- the **interfacing side** will also be a focus of the Rainbow activities. As explained above, we will be interested in both the *forward* (human $\rightarrow$ robot) and *backward* (robot $\rightarrow$ human) interfaces. The forward interface will be mainly addressed from the *algorithmic* point of view, i.e., how to map the few degrees of freedom available to a human operator (usually in the order of 3–4) into complex commands for the controlled robot(s). This mapping will typically be mediated by an "AutoPilot" onboard the robot(s) for autonomously assessing if the commands are feasible and, if not, how to least modify them ("**Advanced Sensor-based Control**" axis).
  The backward interface will, instead, mainly consist of a visual/haptic feedback for the operator. Here, we aim at exploiting our expertise in using force cues for informing an operator about the status of the remote robot(s). However, the sole use of classical *grounded* force feedback devices (e.g., the typical force-feedback joysticks) will not be enough due to the different kinds of information that will have to be provided to the operator. In this context, the recent interest in the use of *wearable* haptic interfaces is very interesting and will be investigated in depth (these include, e.g., devices able to provide vibro-tactile information to the fingertips, wrist, or other parts of the body). The main challenges in these activities will be the mechanical conception (and construction) of suitable wearable interfaces for the tasks at hand, and in the generation of force cues for the operator: the force cues will be a (complex) function of the robot state, therefore motivating research in algorithms for mapping the robot state into a few variables (the force cues) ("**Haptics for Robotics Applications**" axis);

- the **evaluation side** that will assess the proposed interfaces with some user studies, or acceptability studies by human subjects. Although this activity **will not** be a main focus of Rainbow (complex user studies are beyond the scope of our core expertise), we will nevertheless devote some efforts into having some reasonable level of user evaluations by applying standard statistical analysis based on psychophysical procedures (e.g., randomized tests and Anova statistical analysis). This will be particularly true for the activities involving the use of smart wheelchairs, which are intended to be used by human users *and* operate inside human crowds. Therefore, we will be interested in gaining some level of understanding of how semi-autonomous robots (a wheelchair in this example) can predict the human intention, and how humans can react to a semi-autonomous mobile robot.

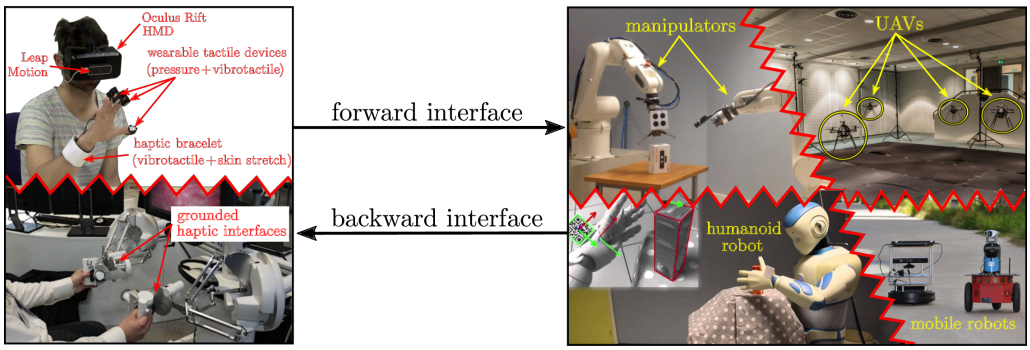


Figure 1. An illustration of the prototypical activities foreseen in Rainbow in which a human operator is in partial (and high-level) control of single/multiple complex robots performing semi-autonomous tasks

Figure 1 depicts in an illustrative way the *prototypical* activities foreseen in Rainbow. On the righthand side, complex robots (dual manipulators, humanoid, single/multiple mobile robots) need to perform some task with high degree of autonomy. On the lefthand side, a human operator gives some high-level commands and receives a visual/haptic feedback aimed at informing her/him at best of the robot status. Again, the main challenges that Rainbow will tackle to address these issues are (in order of relevance): (i) methods and algorithms, mostly based on first-principle modeling and, when possible, on numerical methods for online/reactive trajectory generation, for enabling the robots with high autonomy; (ii) design and implementation of visual/haptic cues for interfacing the human operator with the robots, with a special attention to novel combinations of grounded/ungrounded (wearable) haptic devices; (iii) user and acceptability studies.

3.2. Main Components

Hereafter, a summary description of the four axes of research in Rainbow.

3.2.1. Optimal and Uncertainty-Aware Sensing

Future robots will need to have a large degree of autonomy for, e.g., interpreting the sensory data for accurate estimation of the robot and world state (which can possibly include the human users), and for devising motion plans able to take into account many constraints (actuation, sensor limitations, environment), including also the state estimation accuracy (i.e., how well the robot/environment state can be reconstructed from the sensed data). In this context, we will be particularly interested in (i) devising trajectory optimization strategies able to maximize some norm of the information gain gathered along the trajectory (and with the available sensors). This can be seen as an instance of Active Sensing, with the main focus on *online/reactive* trajectory optimization strategies able to take into account several requirements/constraints (sensing/actuation limitations, noise characteristics). We will also be interested in the coupling between optimal sensing and

concurrent execution of additional tasks (e.g., navigation, manipulation). (ii) Formal methods for guaranteeing the accuracy of localization/state estimation in mobile robotics, mainly exploiting tools from interval analysis. The interest in these methods is their ability to provide possibly conservative but guaranteed accuracy bounds on the best accuracy one can obtain with the given robot/sensor pair, and can thus be used for planning purposes of for system design (choice of the best sensor suite for a given robot/task). (iii) Localization/tracking of objects with poor/unknown or deformable shape, which will be of paramount importance for allowing robots to estimate the state of "complex objects" (e.g., human tissues in medical robotics, elastic materials in manipulation) for controlling its pose/interaction with the objects of interest.

3.2.2. Advanced Sensor-based Control

One of the main competences of the previous Lagadic team has been, generally speaking, the topic of *sensor-based control*, i.e., how to exploit (typically onboard) sensors for controlling the motion of fixed/ground robots. The main emphasis has been in devising ways to directly couple the robot motion with the sensor outputs in order to invert this mapping for driving the robots towards a configuration specified as a desired sensor reading (thus, directly in sensor space). This general idea has been applied to very different contexts: mainly standard vision (from which the Visual Servoing keyword), but also audio, ultrasound imaging, and RGB-D.

Use of sensors for controlling the robot motion will also clearly be a central topic of the Rainbow team too, since the use of (especially onboard) sensing is a main characteristics of any future robotics application (which should typically operate in unstructured environments, and thus mainly rely on its own ability to sense the world). We then naturally aim at making the best out of the previous Lagadic experience in sensor-based control for proposing new advanced ways of exploiting sensed data for, roughly speaking, controlling the motion of a robot. In this respect, we plan to work on the following topics: (i) "direct/dense methods" which try to directly exploit the raw sensory data in computing the control law for positioning/navigation tasks. The advantages of these methods is the need for little data pre-processing which can minimize feature extraction errors and, in general, improve the overall robustness/accuracy (since all the available data is used by the motion controller); (ii) sensor-based interaction with objects of unknown/deformable shapes, for gaining the ability to manipulate, e.g., flexible objects from the acquired sensed data (e.g., controlling online a needle being inserted in a flexible tissue); (iii) sensor-based model predictive control, by developing *online/reactive* trajectory optimization methods able to plan feasible trajectories for robots subjects to sensing/actuation constraints with the possibility of (onboard) sensing for continuously replanning (over some future time horizon) the optimal trajectory. These methods will play an important role when dealing with complex robots affected by complex sensing/actuation constraints, for which pure reactive strategies (as in most of the previous Lagadic works) are not effective. Furthermore, the coupling with the aforementioned optimal sensing will also be considered; (iv) multi-robot decentralised estimation and control, with the aim of devising again sensor-based strategies for groups of multiple robots needing to maintain a formation or perform navigation/manipulation tasks. Here, the challenges come from the need of devising "simple" decentralized and scalable control strategies under the presence of complex sensing constraints (e.g., when using onboard cameras, limited fov, occlusions). Also, the need of locally estimating global quantities (e.g., common frame of reference, global property of the formation such as connectivity or rigidity) will also be a line of active research.

3.2.3. Haptics for Robotics Applications

In the envisaged *shared* cooperation between human users and robots, the typical sensory channel (besides vision) exploited to inform the human users is most often the force/kinesthetic one (in general, the sense of touch and of applied forces to the human hand or limbs). Therefore, a part of our activities will be devoted to study and advance the use of *haptic* cueing algorithms and interfaces for providing a feedback to the users during the execution of some shared task. We will consider: (i) multi-modal haptic cueing for general teleoperation applications, by studying how to convey information through the kinesthetic and cutaneous channels. Indeed, most haptic-enabled applications typically only involve kinesthetic cues, e.g., the forces/torques that can be felt by grasping a force-feedback joystick/device. These cues are very informative about, e.g., preferred/forbidden motion directions, but are also inherently limited in their resolution since the kinesthetic channel can easily become overloaded (when too much information is compressed in a single

cue). In recent years, the arise of novel cutaneous devices able to, e.g., provide vibro-tactile feedback on the fingertips or skin, has proven to be a viable solution to *complement* the classical kinesthetic channel. We will then study how to combine these two sensory modalities for different prototypical application scenarios, e.g., 6-dof teleoperation of manipulator arms, virtual fixtures approaches, and remote manipulation of (possibly deformable) objects; (ii) in the particular context of medical robotics, we plan to address the problem of providing haptic cues for typical medical robotics tasks, such as semi-autonomous needle insertion and robot surgery by exploring the use of kinesthetic feedback for rendering the mechanical properties of the tissues, and vibrotactile feedback for providing with guiding information about pre-planned paths (with the aim of increasing the usability/acceptability of this technology in the medical domain); (iii) finally, in the context of multi-robot control we would like to explore how to use the haptic channel for providing information about the status of *multiple* robots executing a navigation or manipulation task. In this case, the problem is (even more) how to map (or compress) information about many robots into a few haptic cues. We plan to use specialized devices, such as actuated exoskeleton gloves able to provide cues to each fingertip of a human hand, or to resort to "compression" methods inspired by the *hand postural synergies* for providing coordinated cues representative of a few (but complex) motions of the multi-robot group, e.g., coordinated motions (translations/expansions/rotations) or collective grasping/transporting.

3.2.4. Shared Control of Complex Robotics Systems

This final and main research axis will exploit the **methods, algorithms and technologies** developed in the previous axes for realizing applications involving complex semi-autonomous robots operating in complex environments together with human users. The *leitmotiv* is to realize advanced *shared control* paradigms, which essentially aim at blending robot autonomy and user's intervention in an optimal way for exploiting the best of both worlds (robot accuracy/sensing/mobility/strength and human's cognitive capabilities). A common theme will be the issue of where to "draw the line" between robot autonomy and human intervention: obviously, there is no general answer, and any design choice will depend on the particular task at hand and/or on the technological/algorithmic possibilities of the robotic system under consideration.

A *prototypical* envisaged application, exploiting and combining the previous three research axes, is as follows: a complex robot (e.g., a two-arm system, a humanoid robot, a multi-UAV group) needs to operate in an environment exploiting its onboard sensors (in general, vision as the main exteroceptive one) and deal with many constraints (limited actuation, limited sensing, complex kinematics/dynamics, obstacle avoidance, interaction with difficult-to-model entities such as surrounding people, and so on). The robot must then possess a quite large autonomy for interpreting and exploiting the sensed data in order to estimate its own state and the environment one ("**Optimal and Uncertainty-Aware Sensing**" axis), and for planning its motion in order to fulfil the task (e.g., navigation, manipulation) by coping with all the robot/environment constraints. Therefore, advanced control methods able to exploit the sensory data at its most, and able to cope *online* with constraints in an optimal way (by, e.g., continuously replanning and predicting over a future time horizon) will be needed ("**Advanced Sensor-based Control**" axis), with a possible (and interesting) coupling with the sensing part for optimizing, at the same time, the state estimation process. Finally, a human operator will typically be in charge of providing high-level commands (e.g., where to go, what to look at, what to grasp and where) that will then be autonomously executed by the robot, with possible local modifications because of the various (local) constraints. At the same time, the operator will also receive *online* visual-force cues informative of, in general, how well her/his commands are executed and if the robot would prefer or suggest other plans (because of the local constraints that are not of the operator's concern). This information will have to be visually and haptically rendered with an optimal combination of cues that will depend on the particular application ("**Haptics for Robotics Applications**" axis).

RITS Project-Team

3. Research Program

3.1. Vehicle guidance and autonomous navigation

Participants: Mohammad Abualhoul, Pranav Agarwal, Said Alexander Alvarado Marin, Sylia Baraka, Pierre de Beaucorps, Fares Bessam, Pierre Bourre, Raoul de Charette, Carlos Flores, Farouk Ghallabi, Manuel Gonzalez, Maximilian Jaritz, Manohar Kv, Imane Mahtout, Kathia Melbouci, Kaouther Messaoud, Fawzi Nashashibi, Fabio Pizzati, Renaud Poncelet, Danut Ovidiu Pop, Luis Roldao, Anne Verroust-Blondet, Leonardo Ward, Itheri Yahiaoui.

There are three basic ways to improve the safety of road vehicles and these ways are all of interest to the project-team. The first way is to assist the driver by giving him better information and warning. The second way is to take over the control of the vehicle in case of mistakes such as inattention or wrong command. The third way is to completely remove the driver from the control loop.

All three approaches rely on information processing. Only the last two involve the control of the vehicle with actions on the actuators, which are the engine power, the brakes and the steering. The research proposed by the project-team is focused on the following elements:

- perception of the environment,
- planning of the actions,
- real-time control.

3.1.1. Perception of the road environment

Participants: Raoul de Charette, Maximilian Jaritz, Farouk Ghallabi, Manohar Kv, Kaouther Messaoud, Fawzi Nashashibi, Fabio Pizzati, Danut Ovidiu Pop, Luis Roldao, Anne Verroust-Blondet, Itheri Yahiaoui.

Either for driver assistance or for fully automated guided vehicle purposes, the first step of any robotic system is to perceive the environment in order to assess the situation around itself. Proprioceptive sensors (accelerometer, gyrometer,...) provide information about the vehicle by itself such as its velocity or lateral acceleration. On the other hand, exteroceptive sensors, such as video camera, laser or GPS devices, provide information about the environment surrounding the vehicle or its localization. Obviously, fusion of data with various other sensors is also a focus of the research.
The following topics are already validated or under development in our team:

- relative ego-localization with respect to the infrastructure, i.e. lateral positioning on the road can be obtained by mean of vision (lane markings) and the fusion with other devices (e.g. GPS);
- global ego-localization by considering GPS measurement and proprioceptive information, even in case of GPS outage;
- road detection by using lane marking detection and navigable free space;
- detection and localization of the surrounding obstacles (vehicles, pedestrians, animals, objects on roads, etc.) and determination of their behavior can be obtained by the fusion of vision, laser or radar based data processing;
- simultaneous localization and mapping as well as mobile object tracking using laser-based and stereovision-based (SLAMMOT) algorithms.

Scene understanding is a large perception problem. In this research axis we have decided to use only computer vision as cameras have evolved very quickly and can now provide much more precise sensing of the scene, and even depth information. Two types of hardware setups were used, namely: monocular vision or stereo vision to retrieve depth information which allow extracting geometry information.

We have initiated several works:

- estimation of the ego motion using monocular scene flow. Although in the state of the art most of the algorithms use a stereo setup, researches were conducted to estimate the ego-motion using a novel approach with a strong assumption.
- bad weather conditions evaluations. Most often all computer vision algorithms work under a transparent atmosphere assumption which assumption is incorrect in the case of bad weather (rain, snow, hail, fog, etc.). In these situations the light ray are disrupted by the particles in suspension, producing light attenuation, reflection, refraction that alter the image processing.
- deep learning for object recognition. New works are being initiated in our team to develop deep learning recognition in the context of heterogeneous data.
- deep learning for vehicle motion prediction.

3.1.2. Planning and executing vehicle actions

Participants: Pierre de Beaucorps, Carlos Flores, Imane Mahtout, Fawzi Nashashibi, Renaud Poncelet, Anne Verroust-Blondet.

From the understanding of the environment, thanks to augmented perception, we have either to warn the driver to help him in the control of his vehicle, or to take control in case of a driverless vehicle. In simple situations, the planning might also be quite simple, but in the most complex situations we want to explore, the planning must involve complex algorithms dealing with the trajectories of the vehicle and its surroundings (which might involve other vehicles and/or fixed or moving obstacles). In the case of fully automated vehicles, the perception will involve some map building of the environment and obstacles, and the planning will involve partial planning with periodical recomputation to reach the long term goal. In this case, with vehicle to vehicle communications, what we want to explore is the possibility to establish a negotiation protocol in order to coordinate nearby vehicles (what humans usually do by using driving rules, common sense and/or non verbal communication). Until now, we have been focusing on the generation of geometric trajectories as a result of a maneuver selection process using grid-based rating technique or fuzzy technique. For high speed vehicles, Partial Motion Planning techniques we tested, revealed their limitations because of the computational cost. The use of quintic polynomials we designed, allowed us to elaborate trajectories with different dynamics adapted to the driver profile. These trajectories have been implemented and validated in the JointSystem demonstrator of the German Aerospace Center (DLR) used in the European project HAVEit, as well as in RITS's electrical vehicle prototype used in the French project ABV. HAVEit was also the opportunity for RITS to take in charge the implementation of the Co-Pilot system which processes perception data in order to elaborate the high level command for the actuators. These trajectories were also validated on RITS's cybercars. However, for the low speed cybercars that have pre-defined itineraries and basic maneuvers, it was necessary to develop a more adapted planning and control system. Therefore, we have developed a nonlinear adaptive control for automated overtaking maneuver using quadratic polynomials and Lyapunov function candidate and taking into account the vehicles kinematics. For the global mobility systems we are developing, the control of the vehicles includes also advanced platooning, automated parking, automated docking, etc. For each functionality a dedicated control algorithm was designed (see publication of previous years).

3.2. Mobile wireless communications for vehicular networks

Participants: Gérard Le Lann, Mohammad Abualhoul, Fawzi Nashashibi.

Wireless communications are expected to play an essential role in ensuring road safety, road efficiency, and driving comfort o. Road safety applications often require relatively short response time and reliable information exchange between neighboring vehicles and road-side units in any road density condition. Because of the performance of the existing radio communications technology largely degrades with the increase of the traffic density, the challenge of designing wireless communications solution suitable for safety applications is enabling reliable communications in highly dense scenarios.

To investigate this open problem and trade-off situations, RITS has been working on medium access control design for the IEEE 802.11p radio communication and the deployment of supportive solutions such as visible light communications and testing the use-cases for extreme traffic conditions and highly dense scenarios. The works have been carried out considering the vehicle behavior such as autonomous and connected vehicles merging, sharing, and convoy forming as platoon scenarios with considering the hard-safety requirements.

Unlike many of the road safety applications, the applications regarding road efficiency and comfort of road users, often require connectivity to the Internet. Based on our expertise in both Internet-based communications in the mobility context and in ITS, we are investigating the use of IPv6 (Internet Protocol version 6 which is going to replace the current version, IPv4, IoT) for vehicular communications, in a combined architecture supporting both V2V and V2I.

Communication contributions at RITS team have been working on channel modeling for both radio and visible light communications, and design of communications mechanisms, especially for security, service discovery, multicast, and Geo-Cast message delivery, and access point selection.

RITS-team has one of the latest certified standard communication hardware and tools supported by the partnership with the YoGoKo Company. All platforms (connected and autonomous vehicles) are equipped with state-of-art communication units On-Board-Units (OBU), where the Rocquencourt site equipped with two stationary Road-Side-Units (RSU) enabling all kind of tests and projects requirements

Below follows a more detailed description of the related research issues.

3.2.1. Regulation study for interoperability tests for cooperative driving

Participants: Mohammad Abualhoul, Fawzi Nashashibi.

The technological advances of autonomous and connected road vehicles have been shown an accelerating pace in the recent years. On the other hand, the regulations for autonomous, or driverless, road vehicles across Europe still deserve much attention and discussion

Therefore, RITS-Inria team plays a key element in one of the European demonstration-based projects (AUTOC-ITS), which aims to contribute to the regulation study for interoperability in the adoption of autonomous driving in European urban nodes. The regulation study done by RITS team and project partners meant to conduct a deployment of Cooperative Intelligent Transport Systems (C-ITS) in Europe by enhancing interoperability for autonomous vehicles [29]. The project activities and RITS contributions will also boost the role of C-ITS as the primary catalyst for any future implementation of autonomous driving scenarios in Europe. The final demonstration of different European partners will require the implementation and preparations of three pilots sites in three major European cities: Paris, Madrid, and Lisbon. Pilot locations in these major cities are chosen to be located along the European Atlantic Corridor for interoperability evaluation.

RITS-Inria is coordinating the French contribution by evaluating the deployment of C-ITS services in the A13-Paris, which belongs to the French part of the Atlantic Corridor.

Team Core contributions:

- Provide up to date feedback to contribute to the present EU and international regulations on autonomous vehicles.
- Build and evaluate the pilots experimentally by deploying fully autonomous vehicles and a Cooperative Intelligent Transport Systems (C-ITS).
- Define and evaluate a safety autonomous driving services, such as:
  - Roadworks warning.
  - Weather conditions.
  - Other hazardous notifications.
- Define and perform communication interoperability tests between deferent partners for different scenarios, messaging and hardware to ensure the compatibility in using the IEEE 802.11p standard.
- Study the extension of the results on large-scale deployment in other European countries.
- Contribute to the European standards organizations such as C-Roads, C-ITS platforms.

AUTOC-ITS project brings the road authorities from France, Span, and Portugal (DGT, ANSR, SANEF) and C-ITS experts from research institutes and universities (Inria, INDRA, UPM, UC, IPN) to carry out a cooperative work and contributes to the C-ITS Platform by bringing answers to the field of automation driving.

3.2.2. V2X radio communications for road safety applications

Participants: Mohammad Abualhoul, Fawzi Nashashibi.

The development work and generating proper components to facilitate communication requirements and to be deployed in different projects scenarios is one of the main ongoing activities by all RITS team members.

There are continuous activities on both theoretical modeling and experimental evaluation of the radio channel characteristics in vehicular networks, especially the radio quality, channel congestion, load allocations, congestion, and bandwidth availability.

Based on our previous expertise and studies, we develop mechanisms for efficient and reliable V2X communications, access point selection, handover algorithms which are especially dedicated to road safety and autonomous driving applications.

3.2.3. Cyberphysical constructs and mobile communications for fully automated networked vehicles

Participant: Gérard Le Lann.

Intelligent vehicular networks (IVNs) are constituents of ITS. IVNs range from platoons with a lead vehicle piloted by a human driver to fully ad-hoc vehicular networks, a.k.a. VANETs, comprising autonomous/automated vehicles. Safety issues in IVNs appear to be the least studied in the ITS domain. The focus of our work is on safety-critical (SC) scenarios, where accidents and fatalities inevitably occur when such scenarios are not handled correctly. In addition to on-board robotics, inter-vehicular radio communications have been considered for achieving safety properties. Since both technologies have known intrinsic limitations (in addition to possibly experiencing temporary or permanent failures), using them redundantly is mandatory for meeting safety regulations. Redundancy is a fundamental design principle in every SC cyber-physical domain, such as, e.g., air transportation. (Optics-based inter-vehicular communications may also be part of such redundant constructs.) The focus of our on-going work is on safety-critical (SC) communications. We consider IVNs on main roads and highways, which are settings where velocities can be very high, thus exacerbating safety problems acceptable delays in the cyber space, and response times in the physical space, shall be very small. Human lives being at stake, such delays and response times must have strict (non-stochastic) upper bounds under worst-case conditions (vehicular density, concurrency and failures). Consequently, we are led to look for deterministic solutions.

Rationale

In the current ITS literature, the term *safety* is used without being given a precise definition. That must be corrected. In our case, a fundamental open question is: what is the exact meaning of *SC communications*? We have devised a definition, referred to as space-time bounds acceptability (STBA) requirements. For any given problem related to SC communications, those STBA requirements serve as yardsticks for distinguishing acceptable solutions from unacceptable ones with respect to safety. In conformance with the above, STBA requirements rest on the following worst-case upper bounds: λ for channel access delays, and Δ for distributed inter-vehicular coordination (message dissemination, distributed agreement).

Via discussions with foreign colleagues, notably those active in the IEEE 802 Committee, we have comforted our early diagnosis regarding existing standards for V2V/V2I/V2X communications, such as IEEE 802.11p and ETSI ITS-G5: they are totally inappropriate regarding SC communications. A major flaw is the choice of CSMA/CA as the MAC-level protocol. Obviously, there cannot be such bounds as λ and Δ with CSMA/CA. Another flaw is the choice of medium-range omnidirectional communications, radio range in the order of 250 m, and interference range in the order of 400 m. Stochastic delays achievable with existing standards are just unacceptable in moderate/worst-case contention conditions. Consider the following setting, not uncommon in many countries: a highway, 3 lanes each direction, dense traffic, i.e. 1 vehicle per 12.5 m. A simple

calculation leads to the following result: any vehicle may experience (destructive) interferences from up to 384 vehicles. Even if one assumes some reasonable communications activity ratio, say 25%, one finds that up to 96 vehicles may be contending for channel access. Under such conditions, MAC-level delays and string-wide dissemination/agreement delays achieved by current standards fail to meet the STBA requirements by huge margins.

Reliance on V2I communications via terrestrial infrastructures and nodes, such as road-side units or WiFi hotspots, rather than direct V2V communications, can only lead to poorer results. First, reachability is not guaranteed: hazardous conditions may develop anywhere anytime, far away from a terrestrial node. Second, mixing SC communications and ordinary communications within terrestrial nodes is a violation of the very fundamental segregation principle: SC communications and processing shall be isolated from ordinary communications and processing. Third, security: it is very easy to jam or to spy on a terrestrial node; moreover, terrestrial nodes may be used for launching all sorts of attacks, man-in-the-middle attacks for example. Fourth, delays can only get worse than with direct V2V communications, since transiting via a node inevitably introduces additional latencies. Fifth, the delivery of every SC message must be acknowledged, which exacerbates the latency problems. Sixth, availability: what happens when a terrestrial node fails?

Trying to tweak existing standards for achieving SC communications is vain. That is also unjustified. Clearly, medium-range omnidirectional communications are unjustified for the handling of SC scenarios. By definition, accidents can only involve vehicles that are very close to each other. Therefore, short-range directional communications suffice. The obvious conclusion is that novel protocols and inter-vehicular coordination algorithms based on short-range direct V2V communications are needed. It is mandatory to check whether these novel solutions meet the STBA requirements. Future standards specifically aimed at SC communications in IVNs may emerge from such solutions.

Naming and privacy

Additionally, we are exploring the (re)naming problem as it arises in IVNs. Source and destination names appear in messages exchanged among vehicles. Most often, names are IP addresses or MAC addresses (plate numbers shall not be used for privacy reasons). A vehicle which intends to communicate with some vehicle, denoted V here, must know which name *name(V)* to use in order to reach/designate V. Existing solutions are based on multicasting/broadcasting existential messages, whereby every vehicle publicizes its existence (name and geolocation), either upon request (replying to a Geocast) or spontaneously (periodic beaconing). These solutions have severe drawbacks. First, they contribute to overloading communication channels (leading to unacceptably high worst-case delays). Second, they amount to breaching privacy voluntarily. Why should vehicles reveal their existence and their time dependent geolocations, making tracing and spying much easier? Novel solutions are needed. They shall be such that:

- At any time, a vehicle can assign itself a name that is unique within a geographical zone centered on that vehicle (no third-party involved),
- No linkage may exist between a name and those identifiers (plate numbers, IP/MAC addresses, etc.) proper to a vehicle,
- Different (unique) names can be computed at different times by a vehicle (names can be short-lived or long-lived),
- *name(V)* at UTC time t is revealed only to those vehicles sufficiently close to V at time t, notably those which may collide with V.

We have solved the (re)naming problem in string/cohort formations [34]. Ranks (unique integers in any given string/cohort) are privacy-preserving names, easily computed by every member of a string, in the presence of string membership changes (new vehicles join in, members leave). That problem is open when considering arbitrary clusters of vehicles/strings encompassing multiple lanes.

3.3. Probabilistic modeling for large transportation systems

Participants: Guy Fayolle, Jean-Marc Lasgouttes.

This activity concerns the modeling of random systems related to ITS, through the identification and development of solutions based on probabilistic methods and more specifically through the exploration of links between large random systems and statistical physics. Traffic modeling is a very fertile area of application for this approach, both for macroscopic (fleet management [32], traffic prediction) and for microscopic (movement of each vehicle, formation of traffic jams) analysis. When the size or volume of structures grows (leading to the so-called "thermodynamic limit"), we study the quantitative and qualitative (performance, speed, stability, phase transitions, complexity, etc.) features of the system.

In the recent years, several directions have been explored.

3.3.1. Traffic reconstruction

Large random systems are a natural part of macroscopic studies of traffic, where several models from statistical physics can be fruitfully employed. One example is fleet management, where one main issue is to find optimal ways of reallocating unused vehicles: it has been shown that Coulombian potentials might be an efficient tool to drive the flow of vehicles. Another case deals with the prediction of traffic conditions, when the data comes from probe vehicles instead of static sensors.

While the widely-used macroscopic traffic flow models are well adapted to highway traffic, where the distance between junction is long (see for example the work done by the NeCS team in Grenoble), our focus is on a more urban situation, where the graphs are much denser. The approach we are advocating here is model-less, and based on statistical inference rather than fundamental diagrams of road segments. Using the Ising model or even a Gaussian Random Markov Field, together with the very popular Belief Propagation (BP) algorithm, we have been able to show how real-time data can be used for traffic prediction and reconstruction (in the space-time domain).

This new use of BP algorithm raises some theoretical questions about the ways the make the belief propagation algorithm more efficient:

- find the best way to inject real-valued data in an Ising model with binary variables [36];
- build macroscopic variables that measure the overall state of the underlying graph, in order to improve the local propagation of information [33];
- make the underlying model as sparse as possible, in order to improve BP convergence and quality [35].

3.3.2. Exclusion processes for road traffic modeling

The focus here is on road traffic modeled as a granular flow, in order to analyze the features that can be explained by its random nature. This approach is complementary to macroscopic models of traffic flow (as done for example in the Opale team at Inria), which rely mainly on ODEs and PDEs to describe the traffic as a fluid.

One particular feature of road traffic that is of interest to us is the spontaneous formation of traffic jams. It is known that systems as simple as the Nagel-Schreckenberg model are able to describe traffic jams as an emergent phenomenon due to interaction between vehicles. However, even this simple model cannot be explicitly analyzed and therefore one has to resort to simulation.

One of the simplest solvable (but non trivial) probabilistic models for road traffic is the exclusion process. It lends itself to a number of extensions allowing to tackle some particular features of traffic flows: variable speed of particles, synchronized move of consecutive particles (platooning), use of geometries more complex than plain 1D (cross roads or even fully connected networks), formation and stability of vehicle clusters (vehicles that are close enough to establish an ad-hoc communication system), two-lane roads with overtaking.

The aspect that we have particularly studied is the possibility to let the speed of vehicle evolve with time. To this end, we consider models equivalent to a series of queues where the pair (service rate, number of customers) forms a random walk in the quarter plane $\mathbb{Z}_+^2$.

Having in mind a global project concerning the analysis of complex systems, we also focus on the interplay between discrete and continuous description: in some cases, this recurrent question can be addressed quite rigorously via probabilistic methods.

We have considered in [30] some classes of models dealing with the dynamics of discrete curves subjected to stochastic deformations. It turns out that the problems of interest can be set in terms of interacting exclusion processes, the ultimate goal being to derive hydrodynamic limits after proper scaling. A seemingly new method is proposed, which relies on the analysis of specific partial differential operators, involving variational calculus and functional integration. Starting from a detailed analysis of the Asymmetric Simple Exclusion Process (ASEP) system on the torus $\mathbb{Z}/n\mathbb{Z}$, the arguments a priori work in higher dimensions (ABC, multi-type exclusion processes, etc), leading to systems of coupled partial differential equations of Burgers' type.

3.3.3. Random walks in the quarter plane $\mathbb{Z}_+^2$

This field remains one of the important *violon d'Ingres* in our research activities in stochastic processes, both from theoretical and applied points of view. In particular, it is a building block for models of many communication and transportation systems.

One essential question concerns the computation of stationary measures (when they exist). As for the answer, it has been given by original methods formerly developed in the team (see books and related bibliography). For instance, in the case of small steps (jumps of size one in the interior of $\mathbb{Z}_+^2$), the invariant measure $\{\pi_{i,j}, i,j \geq 0\}$ does satisfy the fundamental functional equation (see [2]):

$$Q(x,y)\pi(x,y) = q(x,y)\pi(x) + \widetilde{q}(x,y)\widetilde{\pi}(y) + \pi_0(x,y). \tag{2}$$

where the unknown generating functions $\pi(x,y), \pi(x), \widetilde{\pi}(y), \pi_0(x,y)$ are sought to be analytic in the region $\{(x,y) \in \mathbb{C}^2 : |x| < 1, |y| < 1\}$, and continuous on their respective boundaries.

The given function $Q(x,y) = \sum_{i,j} p_{i,j} x^i y^j - 1$, where the sum runs over the possible jumps of the walk inside $\mathbb{Z}_+^2$, is often referred to as the *kernel*. Then it has been shown that equation (1) can be solved by reduction to a boundary-value problem of Riemann-Hilbert type. This method has been the source of numerous and fruitful developments. Some recent and ongoing works have been dealing with the following matters.

- *Group of the random walk*. In several studies, it has been noticed that the so-called *group of the walk* governs the behavior of a number of quantities, in particular through its *order*, which is always even. In the case of small jumps, the algebraic curve R defined by $\{Q(x,y) = 0\}$ is either of *genus* 0 (the sphere) or 1 (the torus). In [Fayolle-2011a], when the drift of the random walk is equal to 0 (and then so is the genus), an effective criterion gives the *order* of the group. More generally, it is also proved that whenever the genus is 0, this order is infinite, except precisely for the zero drift case, where finiteness is quite possible. When the *genus* is 1, the situation is more difficult. Recently [31], a criterion has been found in terms of a determinant of order 3 or 4, depending on the arity of the group.
- *Nature of the counting generating functions*. Enumeration of planar lattice walks is a classical topic in combinatorics. For a given set of allowed jumps (or steps), it is a matter of counting the number of paths starting from some point and ending at some arbitrary point in a given time, and possibly restricted to some regions of the plane. A first basic and natural question arises: how many such paths exist? A second question concerns the nature of the associated counting generating functions (CGF): are they rational, algebraic, holonomic (or D-finite, i.e. solution of a linear differential equation with polynomial coefficients)?

  Let $f(i,j,k)$ denote the number of paths in $\mathbb{Z}_+^2$ starting from $(0,0)$ and ending at (i,j) at time k. Then the corresponding CGF

$$F(x,y,z) = \sum_{i,j,k \geq 0} f(i,j,k) x^i y^j z^k \tag{3}$$

satisfies the functional equation

$$K(x,y)F(x,y,z) = c(x)F(x,0,z) + \widetilde{c}(y)F(0,y,z) + c_0(x,y), \tag{4}$$

where z is considered as a time-parameter. Clearly, equations (2) and (1) are of the same nature, and answers to the above questions have been given in [Fayolle-2010].

- *Some exact asymptotics in the counting of walks in* $\mathbb{Z}_+^2$. A new and uniform approach has been proposed about the following problem: *What is the asymptotic behavior, as their length goes to infinity, of the number of walks ending at some given point or domain (for instance one axis)?* The method in [Fayolle-2012] works for *both* finite or infinite groups, and for walks not necessarily restricted to excursions.

3.3.4. Simulation for urban mobility

We have worked on various simulation tools to study and evaluate the performance of different transportation modes covering an entire urban area.

- Discrete event simulation for collective taxis, a public transportation system with a service quality comparable with that of conventional taxis.
- Discrete event simulation a system of self-service cars that can reconfigure themselves into shuttles, therefore creating a multimodal public transportation system; this second simulator is intended to become a generic tool for multimodal transportation.
- Joint microscopic simulation of mobility and communication, necessary for investigation of cooperative platoons performance.

These two programs use a technique allowing to run simulations in batch mode and analyze the dynamics of the system afterward.

LINKMEDIA Project-Team

3. Research Program

3.1. Scientific background

LINKMEDIA is de facto a multidisciplinary research team in order to gather the multiple skills needed to enable humans to gain insight into extremely large collections of multimedia material. It is *multimedia data* which is at the core of the team and which drives the design of our scientific contributions, backed-up with solid experimental validations. *Multimedia data*, again, is the rationale for selecting problems, applicative fields and partners.

Our activities therefore include studying the following scientific fields:

- multimedia: content-based analysis; multimodal processing and fusion; multimedia applications;
- computer vision: compact description of images; object and event detection;
- machine learning: deep architectures; structured learning; adversarial learning;
- natural language processing: topic segmentation; information extraction;
- information retrieval: high-dimensional indexing; approximate k-nn search; embeddings;
- data mining: time series mining; knowledge extraction.

3.2. Workplan

Overall, LINKMEDIA follows two main directions of research that are (i) extracting and representing information from the documents in collections, from the relationships between the documents and from what user build from these documents, and (ii) facilitating the access to documents and to the information that has been elaborated from their processing.

3.2.1. Research Direction 1: Extracting and Representing Information

LINKMEDIA follows several research tracks for *extracting* knowledge from the collections and *representing* that knowledge to facilitate users acquiring gradual, long term, constructive insights. Automatically processing documents makes it crucial to consider the accountability of the algorithms, as well as understanding when and why algorithms make errors, and possibly invent techniques that compensate or reduce the impact of errors. It also includes dealing with malicious adversaries carefully manipulating the data in order to compromise the whole knowledge extraction effort. In other words, LINKMEDIA also investigates various aspects related to the *security* of the algorithms analyzing multimedia material for knowledge extraction and representation.

Knowledge is not solely extracted by algorithms, but also by humans as they gradually get insight. This human knowledge can be materialized in computer-friendly formats, allowing algorithms to use this knowledge. For example, humans can create or update ontologies and knowledge bases that are in relation with a particular collection, they can manually label specific data samples to facilitate their disambiguation, they can manually correct errors, etc. In turn, knowledge provided by humans may help algorithms to then better process the data collections, which provides higher quality knowledge to humans, which in turn can provide some better feedback to the system, and so on. This virtuous cycle where algorithms and humans cooperate in order to make the most of multimedia collections requires specific support and techniques, as detailed below.

3.2.1.1. Machine Learning for Multimedia Material.

Many approaches are used to extract relevant information from multimedia material, ranging from very low-level to higher-level descriptions (classes, captions, ...). That diversity of information is produced by algorithms that have varying degrees of supervision. Lately, fully supervised approaches based on deep learning proved to outperform most older techniques. This is particularly true for the latest developments of Recurrent Neural Networkds (RNN, such as LSTMs) or convolutional neural network (CNNs) for images that reach excellent performance [62]. LINKMEDIA contributes to advancing the state of the art in computing representations for multimedia material by investigating the topics listed below. Some of them go beyond the very processing of multimedia material as they also question the fundamentals of machine learning procedures when applied to multimedia.

- *Learning from few samples/weak supervisions.* CNNs and RNNs need large collections of carefully annotated data. They are not fitted for analyzing datasets where few examples per category are available or only cheap image-level labels are provided. LINKMEDIA investigates low-shot, semi-supervised and weakly supervised learning processes: Augmenting scarce training data by automatically propagating labels [65], or transferring what was learned on few very well annotated samples to allow the precise processing of poorly annotated data [74]. Note that this context also applies to the processing of heritage collections (paintings, illuminated manuscripts, ...) that strongly differ from contemporary natural images. Not only annotations are scarce, but the learning processes must cope with material departing from what standard CNNs deal with, as classes such as "planes", "cars", etc, are irrelevant in this case.
- *Ubiquitous Training.* NN (CNNs, LSTMs) are mainstream for producing representations suited for high-quality classification. Their training phase is ubiquitous because the same representations can be used for tasks that go beyond classification, such as retrieval, few-shot, meta- and incremental learning, all boiling down to some form of metric learning. We demonstrated that this ubiquitous training is relatively simpler [65] yet as powerful as ad-hoc strategies fitting specific tasks [79]. We study the properties and the limitations of this ubiquitous training by casting metric learning as a classification problem.
- *Beyond static learning*. Multimedia collections are by nature continuously growing, and ML processes must adapt. It is not conceivable to re-train a full new model at every change, but rather to support continuous training and/or allowing categories to evolve as the time goes by. New classes may be defined from only very few samples, which links this need for dynamicity to the low-shot learning problem discussed here. Furthermore, active learning strategies determining which is the next sample to use to best improve classification must be considered to alleviate the annotation cost and the re-training process [69]. Eventually, the learning process may need to manage an extremely large number of classes, up to millions. In this case, there is a unique opportunity of blending the expertise of LINKMEDIA on large scale indexing and retrieval with deep learning. Base classes can either be "summarized" e.g. as a multi-modal distribution, or their entire training set can be made accessible as an external associative memory [86].
- *Learning and lightweight architectures.* Multimedia is everywhere, it can be captured and processed on the mobile devices of users. It is necessary to study the design of lightweight ML architectures for mobile and embedded vision applications. Inspired by [90], we study the savings from quantizing hyper-parameters, pruning connections or other approximations, observing the trade-off between the footprint of the learning and the quality of the inference. Once strategy of choice is progressive learning which early aborts when confident enough [70].
- *Multimodal embeddings.* We pursue pioneering work of LINKMEDIA on multimodal embedding, i.e., representing multiple modalities or information sources in a single embedded space [83], [85], [84]. Two main directions are explored: exploiting adversarial architectures (GANs) for embedding via translation from one modality to another, extending initial work in [84] to highly heterogeneous content; combining and constraining word and RDF graph embeddings to facilitate entity linking and explanation of lexical co-occurrences [81].

- *Accountability of ML processes.* ML processes achieve excellent results but it is mandatory to verify that accuracy results from having determined an adequate problem representation, and not from being abused by artifacts in the data. LINKMEDIA designs procedures for at least explaining and possibly interpreting and understanding what the models have learned. We consider heat-maps materializing which input (pixels, words) have the most importance in the decisions [77], Taylor decompositions to observe the individual contributions of each relevance scores or estimating LID [47] as a surrogate for accounting for the smoothness of the space.
- *Extracting information.* ML is good at extracting features from multimedia material, facilitating subsequent classification, indexing, or mining procedures. LINKMEDIA designs extraction processes for identifying parts in the images [75], [76], relationships between the various objects that are represented in images [53], learning to localizing objects in images with only weak, image-level supervision [78] or fine-grained semantic information in texts [58]. One technique of choice is to rely on generative adversarial networks (GAN) for learning low-level representations. These representations can e.g. be based on the analysis of density [89], shading, albedo, depth, etc.
- *Learning representations for time evolving multimedia material.* Video and audio are time evolving material, and processing them requests to take their time line into account. In [71], [57] we demonstrated how shapelets can be used to transform time series into time-free high-dimensional vectors, preserving however similarities between time series. Representing time series in a metric space improves clustering, retrieval, indexing, metric learning, semi-supervised learning and many other machine learning related tasks. Research directions include adding localization information to the shapelets, fine-tuning them to best fit the task in which they are used as well as designing hierarchical representations.

3.2.1.2. Adversarial Machine Learning.

Systems based on ML take more and more decisions on our behalf, and maliciously influencing these decisions by crafting adversarial multimedia material is a potential source of dangers: a small amount of carefully crafted noise imperceptibly added to images corrupts classification and/or recognition. This can naturally impact the insight users get on the multimedia collection they work with, leading to taking erroneous decisions e.g.

This adversarial phenomenon is not particular to deep learning, and can be observed even when using other ML approaches [52]. Furthermore, it has been demonstrated that adversarial samples generalize very well across classifiers, architectures, training sets. The reasons explaining why such tiny content modifications succeed in producing severe errors are still not well understood.

We are left with little choice: we must gain a better understanding of the weaknesses of ML processes, and in particular of deep learning. We must understand why attacks are possible as well as discover mechanisms protecting ML against adversarial attacks (with a special emphasis on convolutional neural networks). Some initial contributions have started exploring such research directions, mainly focusing on images and computer vision problems. Very little has been done for understanding adversarial ML from a *multimedia* perspective [56].

LINKMEDIA is in a unique position to throw at this problem new perspectives, by experimenting with other modalities, used in isolation one another, as well as experimenting with true multimodal inputs. This is very challenging, and far more complicated and interesting than just observing adversarial ML from a computer vision perspective. No one clearly knows what is at stake with adversarial audio samples, adversarial video sequences, adversarial ASR, adversarial NLP, adversarial OCR, all this being often part of a sophisticated multimedia processing pipeline.

Our ambition is to lead the way for initiating investigations where the full diversity of modalities we are used to work with in multimedia are considered from a perspective of adversarial attacks and defenses, both at learning and test time. In addition to what is described above, and in order to trust the multimedia material we analyze and/or the algorithms that are at play, LINKMEDIA investigates the following topics:

- *Beyond classification.* Most contributions in relation with adversarial ML focus on classification tasks. We started investigating the impact of adversarial techniques on more diverse tasks such as

retrieval [46]. This problem is related to the very nature of euclidean spaces where distances and neighborhoods can all be altered. Designing defensive mechanisms is a natural companion work.

- *Detecting false information.* We carry-on with earlier pioneering work of LINKMEDIA on false information detection in social media. Unlike traditional approaches in image forensics [60], we build on our expertise in content-based information retrieval to take advantage of the contextual information available in databases or on the web to identify out-of-context use of text or images which contributed to creating a false information [72].
- *Deep fakes.* Progress in deep ML and GANs allow systems to generate realistic images and are able to craft audio and video of existing people saying or doing things they never said or did [68]. Gaining in sophistication, these machine learning-based "deep fakes" will eventually be almost indistinguishable from real documents, making their detection/rebutting very hard. LINKMEDIA develops deep learning based counter-measures to identify such modern forgeries. We also carry on with making use of external data in a provenance filtering perspective [91] in order to debunk such deep fakes.
- *Distributions, frontiers, smoothness, outliers.* Many factors that can possibly explain the adversarial nature of some samples are in relation with their distribution in space which strongly differs from the distribution of natural, genuine, non adversarial samples. We are investigating the use of various information theoretical tools that facilitate observing distributions, how they differ, how far adversarial samples are from benign manifolds, how smooth is the feature space, etc. In addition, we are designing original adversarial attacks and develop detection and curating mechanisms [47].

3.2.1.3. Multimedia Knowledge Extraction.

Information obtained from collections via computer ran processes is not the only thing that needs to be represented. Humans are in the loop, and they gradually improve their level of understanding of the content and nature of the multimedia collection. Discovering knowledge and getting insight is involving multiple people across a long period of time, and what each understands, concludes and discovers must be recorded and made available to others. Collaboratively inspecting collections is crucial. Ontologies are an often preferred mechanism for modeling what is inside a collection, but this is probably limitative and narrow.

LINKMEDIA is concerned with making use of existing strategies in relation with ontologies and knowledge bases. In addition, LINKMEDIA uses mechanisms allowing to materialize the knowledge gradually acquired by humans and that might be subsequently used either by other humans or by computers in order to better and more precisely analyze collections. This line of work is instantiated at the core of the iCODA project LINKMEDIA coordinates. We are therefore concerned with:

- *Multimedia analysis and ontologies.* We develop approaches for linking multimedia content to entities in ontologies for text and images, building on results in multimodal embedding to cast entity linking into a nearest neighbor search problem in a high-dimensional joint embedding of content and entities [85]. We also investigate the use of ontological knowledge to facilitate information extraction from content [9].
- *Explainability and accountability in information extraction.* In relation with ontologies and entity linking, we develop innovative approaches to explain statistical relations found in data, in particular lexical or entity co-occurrences in textual data, for example using embeddings constrained with translation properties of RDF knowledge or path-based explanation within RDF graphs. We also work on confidence measures in entity linking and information extraction, studying how the notions of confidence and information source can be accounted for in knowledge basis and used in human-centric collaborative exploration of collections.
- *Dynamic evolution of models for information extraction.* In interactive exploration and information extraction, e.g., on cultural or educational material, knowledge progressively evolves as the process goes on, requiring on-the-fly design of new models for content-based information extractors from very few examples, as well as continuous adaptation of the models. Combining in a seamless way low-shot, active and incremental learning techniques is a key issue that we investigate to enable this dynamic mechanisms on selected applications.

3.2.1.4. Research Direction 2: Accessing Information

LINKMEDIA centers its activities on enabling humans to make good use of vast multimedia collections. This material takes all its cultural and economic value, all its artistic wonder when it can be accessed, watched, searched, browsed, visualized, summarized, classified, shared, ... This allows users to fully enjoy the incalculable richness of the collections. It also makes it possible for companies to create business rooted in this multimedia material.

Accessing the multimedia data that is inside a collection is complicated by the various type of data, their volume, their length, etc. But it is even more complicated to access the information that is not materialized in documents, such as the relationships between parts of different documents that however share some similarity. LINKMEDIA in its first four years of existence established itself as one of the leading teams in the field of multimedia analytics, contributing to the establishment of a dedicated community (refer to the various special sessions we organized with MMM, the iCODA and the LIMAH projects, as well as [66], [67], [63]).

Overall, facilitating the access to the multimedia material, to the relevant information and the corresponding knowledge asks for algorithms that efficiently *search* collections in order to identify the elements of collections or of the acquired knowledge that are matching a query, or that efficiently allow *navigating* the collections or the acquired knowledge. Navigation is likely facilitated if techniques are able to handle information and knowledge according to hierarchical perspectives, that is, allow to reveal data according to various levels of details. Aggregating or *summarizing* multimedia elements is not trivial.

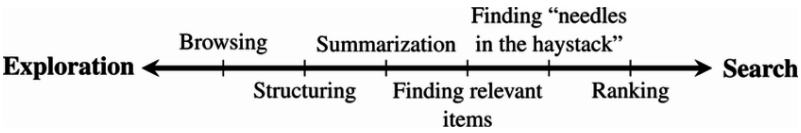


Figure 1. Exploration-search axis with example tasks

Three topics are therefore in relation with this second research direction. LINKMEDIA tackles the issues in relation to searching, to navigating and to summarizing multimedia information. Information needs when discovering the content of a multimedia collection can be conveniently mapped to the exploration-search axis, as first proposed by Zahálka and Worring in [88], and illustrated by Figure 1 where expert users typically work near the right end because their tasks involve precise queries probing search engines. In contrast, lay-users start near the exploration end of the axis. Overall, users may alternate searches and explorations by going back and forth along the axis. The underlying model and system must therefore be highly dynamic, support interactions with the users and propose means for easy refinements. LINKMEDIA contributes to advancing

the state of the art in searching operations, in navigating operations (also referred to as browsing), and in summarizing operations.

3.2.1.4.1. Searching.

Search engines must run similarity searches very efficiently. High-dimensional indexing techniques therefore play a central role. Yet, recent contributions in ML suggest to revisit indexing in order to adapt to the specific properties of modern features describing contents.

- *Advanced scalable indexing.* High-dimensional indexing is one of the foundations of LINKMEDIA. Modern features extracted from the multimedia material with the most recent ML techniques shall be indexed as well. This, however, poses a series of difficulties due to the dimensionality of these features, their possible sparsity, the complex metrics in use, the task in which they are involved (instance search, k-nn, class prototype identification, manifold search [65], time series retrieval, ...). Furthermore, truly large datasets require involving sketching [50], secondary storage and/or distribution [49], [48], alleviating the explosion of the number of features to consider due to their local nature or other innovative methods [64], all introducing complexities. Last, indexing multimodal embedded spaces poses a new series of challenges.
- *Improving quality.* Scalable indexing techniques are approximate, and what they return typically includes a fair amount of false positives. LINKMEDIA works on improving the quality of the results returned by indexing techniques. Approaches taking into account neighborhoods [59], manifold structures instead of pure distance based similarities [65] must be extended to cope with advanced indexing in order to enhance quality. This includes feature selection based on intrinsic dimensionality estimation [47].
- *Dynamic indexing.* Feature collections grow, and it is not an option to fully reindex from scratch an updated collection. This trivially applies to the features directly extracted from the media items, but also to the base class prototypes that can evolve due to the non-static nature of learning processes. LINKMEDIA will continue investigating what is at stake when designing dynamic indexing strategies.

3.2.1.4.2. Navigating.

Navigating a multimedia collection is very central to its understanding. It differs from searching as navigation is not driven by any specific query. Rather, it is mostly driven by the relationships that various documents have one another. Relationships are supported by the links between documents and/or parts of documents. Links rely on semantic similarity, depicting the fact that two documents share information on the same topic. But other aspects than semantics are also at stake, e.g., time with the dates of creation of the documents or geography with mentions or appearance in documents of some geographical landmarks or with geo-tagged data.

In multimedia collections, links can be either implicit or explicit, the latter being much easier to use for navigation. An example of an implicit link can be the name of someone existing in several different news articles; we, as humans, create a mental link between them. In some cases, the computer misses such configurations, leaving such links implicit. Implicit links are subject to human interpretation, hence they are sometimes hard to identify for any automatic analysis process. Implicit links not being materialized, they can therefore hardly be used for navigation or faceted search. Explicit links can typically be seen as hyperlinks, established either by content providers or, more aligned with LINKMEDIA, automatically determined from content analysis. Entity linking (linking content to an entity referenced in a knowledge base) is a good example of the creation of explicit links. Semantic similarity links, as investigated in the LIMAH project and as considered in the search and hyperlinking task at MediaEval and TRECVid, are also prototypical links that can be made explicit for navigation. Pursuing work, we investigate two main issues:

- *Improving multimodal content-based linking.* We exploit achievements in entity linking to go beyond lexical or lexico-visual similarity and to provide semantic links that are easy to interpret for humans; carrying on, we work on link characterization, in search of mechanisms addressing link explainability (i.e., what is the nature of the link), for instance using attention models so as to focus

on the common parts of two documents or using natural language generation; a final topic that we address is that of linking textual content to external data sources in the field of journalism, e.g., leveraging topic models and cue phrases along with a short description of the external sources.

- *Dynamicity and user-adaptation.* One difficulty for explicit link creation is that links are often suited for one particular usage but not for another, thus requiring creating new links for each intended use; whereas link creation cannot be done online because of its computational cost, the alternative is to generate (almost) all possible links and provide users with selection mechanisms enabling personalization and user-adaptation in the exploration process; we design such strategies and investigate their impact on exploration tasks in search of a good trade-off between performance (few high-quality links) and genericity.

3.2.1.4.3. Summarizing.

Multimedia collections contain far too much information to allow any easy comprehension. It is mandatory to have facilities to aggregate and summarize a large body on information into a compact, concise and meaningful representation facilitating getting insight. Current technology suggests that multimedia content aggregation and story-telling are two complementary ways to provide users with such higher-level views. Yet, very few studies already investigated these issues. Recently, video or image captioning [87], [82] have been seen as a way to summarize visual content, opening the door to state-of-the-art multi-document text summarization [61] with text as a pivot modality. Automatic story-telling has been addressed for highly specific types of content, namely TV series [54] and news [73], [80], but still need a leap forward to be mostly automated, e.g., using constraint-based approaches for summarization [51], [80].

Furthermore, not only the original multimedia material has to be summarized, but the knowledge acquired from its analysis is also to summarize. It is important to be able to produce high-level views of the relationships between documents, emphasizing some structural distinguishing qualities. Graphs establishing such relationships need to be constructed at various level of granularity, providing some support for summarizing structural traits.

Summarizing multimedia information poses several scientific challenges that are:

- *Choosing the most relevant multimedia aggregation type*: Taking a multimedia collection into account, a same piece of information can be present in several modalities. The issue of selecting the most suitable one to express a given concept has thus to be considered together with the way to mix the various modalities into an acceptable production. Standard summarization algorithms have to be revisited so that they can handle continuous representation spaces, allowing them to benefit from the various modalities [55].
- *Expressing user's preferences*: Different users may appreciate quite different forms of multimedia summaries, and convenient ways to express their preferences have to be proposed. We for example focus on the opportunities offered by the constraint-based framework.
- *Evaluating multimedia summaries*: Finding criteria to characterize what a good summary is remains challenging, e.g., how to measure the global relevance of a multimodal summary and how to compare information between and across two modalities. We tackle this issue particularly via a collaboration with A. Smeaton at DCU, comparing the automatic measures we will develop to human judgments obtained by crowd-sourcing;
- *Taking into account structuring and dynamicity*: Typed links between multimedia fragments, and hierarchical topical structures of documents obtained via work previously developed within the team are two types of knowledge which have seldom been considered as long as summarization is concerned. Knowing that the event present in a document is causally related to another event described in another document can however modify the ways summarization algorithms have to consider information. Moreover the question of producing coarse-to-fine grain summaries exploiting the topical structure of documents is still an open issue. Summarizing dynamic collections is also challenging and it is one of the questions we consider.

MAGRIT Team

3. Research Program

3.1. Matching and 3D tracking

One of the most basic problems currently limiting AR applications is the registration problem. The objects in the real and virtual worlds must be properly aligned with respect to each other, or the illusion that the two worlds coexist will be compromised.

As a large number of potential AR applications are interactive, real time pose computation is required. Although the registration problem has received a lot of attention in the computer vision community, the problem of real-time registration is still far from being a solved problem, especially for unstructured environments. Ideally, an AR system should work in all environments, without the need to prepare the scene ahead of time, independently of the variations in experimental conditions (lighting, weather condition,...)

For several years, the MAGRIT project has been aiming at developing on-line and marker-less methods for camera pose computation. The main difficulty with on-line tracking is to ensure robustness of the process over time. For off-line processes, robustness is achieved by using spatial and temporal coherence of the considered sequence through move-matching techniques. To get robust open-loop systems, we have investigated various methods, ranging from statistical methods to the use of hybrid camera/sensor systems. Many of these methods are dedicated to piecewise-planar scenes and combine the advantage of move-matching methods and model-based methods. In order to reduce statistical fluctuations in viewpoint computation, which lead to unpleasant jittering or sliding effects, we have also developed model selection techniques which allow us to noticeably improve the visual impression and to reduce drift over time. Another line of research which has been considered in the team to improve the reliability and the robustness of pose algorithms is to combine the camera with another form of sensor in order to compensate for the shortcomings of each technology.

The success of pose computation over time largely depends on the quality of the matching at the initialization stage. Indeed, the current image may be very different from the appearances described in the model both on the geometrical and the photometric sides. Research is thus conducted in the team on the use of probabilistic methods to establish robust correspondences of features. The use of *a contrario* methods has been investigated to achieve this aim [7]. We especially addressed the complex case of matching in scenes with repeated patterns which are common in urban scenes. We are also investigating the problem of matching images taken from very different viewpoints which is central for the re-localization issue in AR. Within the context of a scene model acquired with structure-from-motion techniques, we are currently investigating the use of viewpoint simulation in order to allow successful pose computation even if the considered image is far from the positions used to build the model [15].

Recently, the issue of tracking deformable objects has gained importance in the team. This topic is mainly addressed in the context of medical applications through the design of bio-mechanical models guided by visual features [2]. We have successfully investigated the use of such models in laparoscopy, with a vascularized model of the liver and with a hyper-elastic model for tongue tracking in ultrasound images. However, these results have been obtained so far in relatively controlled environments, with non-pathological cases. When clinical routine applications are to be considered, many parameters and considerations need to be taken into account. Among the problems that need to be addressed are more realistic model representations, the specification of the range of physical parameters and the need to enforce the robustness of the tracking with respect to outliers, which are common in the interventional context.

3.2. Image-based Modeling

Modeling the scene is a fundamental issue in AR for many reasons. First, pose computation algorithms often use a model of the scene or at least some 3D knowledge on the scene. Second, effective AR systems require a

model of the scene to support interactions between the virtual and the real objects such as occlusions, lighting reflections, contacts... in real-time. Unlike pose computation which has to be performed in a sequential way, scene modeling can be considered as an off-line or an on-line problem depending on the requirements of the targeted application. Interactive in-situ modeling techniques have thus been developed with the aim to enable the user to define what is relevant at the time the model is being built during the application. On the other hand, we also proposed off-line multimodal techniques, mainly dedicated to AR medical applications, with the aim of obtaining realistic and possibly dynamic models of organs suitable for real-time simulation [3].

In-situ modeling
In-situ modeling allows a user to directly build a 3D model of his/her surrounding environment and verify the geometry against the physical world in real-time. This is of particular interest when using AR in unprepared environments or building scenes that either have an ephemeral existence (e.g., a film set) or cannot be accessed frequently (e.g., a nuclear power plant). We have especially investigated two systems, one based on the image content only and the other based on multiple data coming from different sensors (camera, inertial measurement unit, laser rangefinder). Both systems use the camera-mouse principle [34] (i.e., interactions are performed by aiming at the scene through a video camera) and both systems have been designed to acquire polygonal textured models, which are particularly useful for camera tracking and object insertion in AR.

Multimodal modeling for real-time simulation
With respect to classical AR applications, AR in medical context differs in the nature and the size of the data which are available: a large amount of multimodal data is acquired on the patient or possibly on the operating room through sensing technologies or various image acquisitions [32]. The challenge is to analyze these data, to extract interesting features, to fuse and to visualize this information in a proper way. Within the MAGRIT team, we address several key problems related to medical augmented environments. Being able to acquire multimodal data which are temporally synchronized and spatially registered is the first difficulty we face when considering medical AR. Another key requirement of AR medical systems is the availability of 3D (+t) models of the organ/patient built from images, to be overlaid onto the users' view of the environment.

Methods for multimodal modeling are strongly dependent on the imaging modalities and the organ specificities. We thus only address a restricted number of medical applications –interventional neuro-radiology, laparoscopic surgery– for which we have a strong expertise and close relationships with motivated clinicians. In these applications, our aim is to produce realistic models and then realistic simulations of the patient to be used for the training of surgeons or the re-education of patients.

One of our main applications is about neuroradiology. For the last 20 years, we have been working in close collaboration with the neuroradiology laboratory (CHRU-University Hospital of Nancy) and GE Healthcare. As several imaging modalities are now available in an intraoperative context (2D and 3D angiography, MRI, ...), our aim is to develop a multi-modality framework to assist therapeutic decision and treatment.

We have mainly been interested in the effective use of a multimodality framework in the treatment of arteriovenous malformations (AVM) and aneurysms in the context of interventional neuroradiology. The goal of interventional gestures is to guide endoscopic tools towards the pathology with the aim to perform embolization of the AVM or to fill the aneurysmal cavity by placing coils. We have proposed and developed multimodality and augmented reality tools which make various image modalities (2D and 3D angiography, fluoroscopic images, MRI, ...) cooperate in order to assist physicians in clinical routine. One of the successes of this collaboration is the implementation of the concept of *augmented fluoroscopy*, which helps the surgeon to guide endoscopic tools towards the pathology. Lately, in cooperation with the team MIMESIS, we have proposed new methods for implicit modeling of the vasculature with the aim of obtaining near real-time simulation of the coil deployment in the aneurysm [3]. These works open the way towards near real-time patient-based simulations of interventional gestures both for training and for planning.

3.3. Parameter estimation

Many problems in computer vision or image analysis can be formulated in terms of parameter estimation from image-based measurements. This is the case of many problems addressed in the team such as pose

computation or image-guided estimation of 3D deformable models. Often traditional robust techniques which take into account the covariance on the measurements are sufficient to achieve reliable parameter estimation. However, depending on their number, their spatial distribution and the uncertainty on these measurements, some problems are very sensitive to noise and there is a considerable interest in considering how parameter estimation could be improved if additional information on the noise were available. Another common problem in our field of research is the need to estimate constitutive parameters of the models, such as (bio)-mechanical parameters for instance. Direct measurement methods are destructive, and elaborating image-based methods is thus highly desirable. Besides designing appropriate estimation algorithms, a fundamental question is to understand what group of parameters under study can be reliably estimated from a given experimental setup.

This line of research is relatively new in the team. One of the challenges is to improve image-based parameter estimation techniques considering sensor noise and specific image formation models. In a collaboration with the Pascal Institute (Clermont Ferrand), metrological performance enhancement for experimental solid mechanics has been addressed through the development of dedicated signal processing methods [6]. In the medical field, specific methods based on an adaptive evolutionary optimization strategy have been designed for estimating respiratory parameters [8]. In the context of designing realistic simulators for neuroradiology, we are now considering how parameters involved in the simulation could be adapted to fit real images.

MORPHEO Project-Team

3. Research Program

3.1. Shape and Appearance Modeling

Standard acquisition platforms, including commercial solutions proposed by companies such as Microsoft, 3dMD or 4DViews, now give access to precise 3D models with geometry, e.g. meshes, and appearance information, e.g. textures. Still, state-of-the-art solutions are limited in many respects: They generally consider limited contexts and close setups with typically at most a few meter side lengths. As a result, many dynamic scenes, even a body running sequence, are still challenging situations; They also seldom exploit time redundancy; Additionally, data driven strategies are yet to be fully investigated in the field. The MORPHEO team builds on the Kinovis platform for data acquisition and has addressed these issues with, in particular, contributions on time integration, in order to increase the resolution for both shapes and appearances, on representations, as well as on exploiting recent machine learning tools when modeling dynamic scenes. Our originality lies, for a large part, in the larger scale of the dynamic scenes we consider as well as in the time super resolution strategy we investigate. Another particularity of our research is a strong experimental foundation with the multiple camera Kinovis platforms.

3.2. Dynamic Shape Vision

Dynamic Shape Vision refers to research themes that consider the motion of dynamic shapes, with e.g. shapes in different poses, or the deformation between different shapes, with e.g. different human bodies. This includes for instance shape tracking, shape registration, all these themes being covered by MORPHEO. While progress has been made over the last decade in this domain, challenges remain, in particular due to the required essential task of shape correspondence that is still difficult to perform robustly. Strategies in this domain can be roughly classified into two categories: (i) data driven approaches that learn shape spaces and estimate shapes and their variations through space parameterizations; (ii) model based approaches that use more or less constrained prior models on shape evolutions, e.g. locally rigid structures, to recover correspondences. The MORPHEO team is substantially involved in the second category that leaves more flexibility for shapes that can be modeled, an important feature with the Kinovis platform. The team is anyway also considering the first category with faces and body under clothes modeling, classes of shapes that are more likely to evolve in spaces with reasonable dimensions. The originality of MORPHEO in this axis is to go beyond static shape poses and to consider also the dynamics of shape over several frames when modeling moving shapes, this in particular with shape tracking, animation and, more recently, face registration.

3.3. Inside Shape Vision

Another research axis is concerned with the ability to perceive inside moving shapes. This is a more recent research theme in the MORPHEO team that has gained importance. It was originally the research associated to the Kinovis platform installed in the Grenoble Hospitals. This platform is equipped with two X-ray cameras and ten color cameras, enabling therefore simultaneous vision of inside and outside shapes. We believe this opens a new domain of investigation at the interface between computer vision and medical imaging. Interesting issues in this domain include the links between the outside surface of a shape and its inner parts, especially with the human body. These links are likely to help understanding and modeling human motions. Until now, numerous dynamic shape models, especially in the computer graphic domain, consist of a surface, typically a mesh, bound to a skeletal structure that is never observed in practice but that help anyway parameterizing human motion. Learning more accurate relationships using observations can therefore significantly impact the domain.

3.4. Shape Animation

3D animation is a crucial part of digital media production with numerous applications, in particular in the game and motion picture industry. Recent evolutions in computer animation consider real videos for both the creation and the animation of characters. The advantage of this strategy is twofold: it reduces the creation cost and increases realism by considering only real data. Furthermore, it allows to create new motions, for real characters, by recombining recorded elementary movements. In addition to enable new media contents to be produced, it also allows to automatically extend moving shape datasets with fully controllable new motions. This ability appears to be of great importance with the recent advent of deep learning techniques and the associated need for large learning datasets. In this research direction, we investigate how to create new dynamic scenes using recorded events.

PERCEPTION Project-Team

3. Research Program

3.1. Audio-Visual Scene Analysis

From 2006 to 2009, R. Horaud was the scientific coordinator of the collaborative European project POP (Perception on Purpose), an interdisciplinary effort to understand visual and auditory perception at the crossroads of several disciplines (computational and biological vision, computational auditory analysis, robotics, and psychophysics). This allowed the PERCEPTION team to launch an interdisciplinary research agenda that has been very active for the last five years. There are very few teams in the world that gather scientific competences spanning computer vision, audio signal processing, machine learning and human-robot interaction. The fusion of several sensorial modalities resides at the heart of the most recent biological theories of perception. Nevertheless, multi-sensor processing is still poorly understood from a computational point of view. In particular and so far, audio-visual fusion has been investigated in the framework of speech processing using close-distance cameras and microphones. The vast majority of these approaches attempt to model the temporal correlation between the auditory signals and the dynamics of lip and facial movements. Our original contribution has been to consider that audio-visual localization and recognition are equally important. We have proposed to take into account the fact that the audio-visual objects of interest live in a three-dimensional physical space and hence we contributed to the emergence of *audio-visual scene analysis* as a scientific topic in its own right. We proposed several novel statistical approaches based on supervised and unsupervised mixture models. The *conjugate mixture model* (CMM) is an unsupervised probabilistic model that allows to cluster observations from different modalities (e.g., vision and audio) living in different mathematical spaces [25], [2]. We thoroughly investigated CMM, provided practical resolution algorithms and studied their convergence properties. We developed several methods for sound localization using two or more microphones [1]. The *Gaussian locally-linear model* (GLLiM) is a partially supervised mixture model that allows to map high-dimensional observations (audio, visual, or concatenations of audio-visual vectors) onto low-dimensional manifolds with a partially known structure [9]. This model is particularly well suited for perception because it encodes both observable and unobservable phenomena. A variant of this model, namely *probabilistic piecewise affine mapping* has also been proposed and successfully applied to the problem of sound-source localization and separation [8]. The European projects HUMAVIPS (2010-2013) coordinated by R. Horaud and EARS (2014-2017), applied audio-visual scene analysis to human-robot interaction.

3.2. Stereoscopic Vision

Stereoscopy is one of the most studied topics in biological and computer vision. Nevertheless, classical approaches of addressing this problem fail to integrate eye/camera vergence. From a geometric point of view, the integration of vergence is difficult because one has to re-estimate the epipolar geometry at every new eye/camera rotation. From an algorithmic point of view, it is not clear how to combine depth maps obtained with different eyes/cameras relative orientations. Therefore, we addressed the more general problem of binocular vision that combines the low-level eye/camera geometry, sensor rotations, and practical algorithms based on global optimization [19], [32]. We studied the link between mathematical and computational approaches to stereo (global optimization and Markov random fields) and the brain plausibility of some of these approaches: indeed, we proposed an original mathematical model for the complex cells in visual-cortex areas V1 and V2 that is based on steering Gaussian filters and that admits simple solutions [20]. This addresses the fundamental issue of how local image structure is represented in the brain/computer and how this structure is used for estimating a dense disparity field. Therefore, the main originality of our work is to address both computational and biological issues within a unifying model of binocular vision. Another equally important problem that still remains to be solved is how to integrate binocular depth maps over time. Recently, we have addressed this problem and proposed a semi-global optimization framework that starts with sparse yet reliable matches and proceeds with propagating them over both space and time. The concept of seed-match propagation has then been extended to TOF-stereo fusion [12].

3.3. Audio Signal Processing

Audio-visual fusion algorithms necessitate that the two modalities are represented in the same mathematical space. Binaural audition allows to extract sound-source localization (SSL) information from the acoustic signals recorded with two microphones. We have developed several methods, that perform sound localization in the temporal and the spectral domains. If a direct path is assumed, one can exploit the *time difference of arrival* (TDOA) between two microphones to recover the position of the sound source with respect to the position of the two microphones. The solution is not unique in this case, the sound source lies onto a 2D manifold. However, if one further assumes that the sound source lies in a horizontal plane, it is then possible to extract the azimuth. We used this approach to predict possible sound locations in order to estimate the direction of a speaker [2]. We also developed a geometric formulation and we showed that with four non-coplanar microphones the azimuth and elevation of a single source can be estimated without ambiguity [1]. We also investigated SSL in the spectral domain. This exploits the filtering effects of the head related transfer function (HRTF): there is a different HRTF for the left and right microphones. The interaural spectral features, namely the ILD (interaural level difference) and IPD (interaural phase difference) can be extracted from the short-time Fourier transforms of the two signals. The sound direction is encoded in these interaural features but it is not clear how to make SSL explicit in this case. We proposed a supervised learning formulation that estimates a mapping from interaural spectral features (ILD and IPD) to source directions using two different setups: audio-motor learning [8] and audio-visual learning [10].

3.4. Visual Reconstruction With Multiple Color and Depth Cameras

For the last decade, one of the most active topics in computer vision has been the visual reconstruction of objects, people, and complex scenes using a multiple-camera setup. The PERCEPTION team has pioneered this field and by 2006 several team members published seminal papers in the field. Recent work has concentrated onto the robustness of the 3D reconstructed data using probabilistic outlier rejection techniques combined with algebraic geometry principles and linear algebra solvers [35]. Subsequently, we proposed to combine 3D representations of shape (meshes) with photometric data [33]. The originality of this work was to represent photometric information as a scalar function over a discrete Riemannian manifold, thus *generalizing image analysis to mesh and graph analysis*. Manifold equivalents of local-structure detectors and descriptors were developed [34]. The outcome of this pioneering work has been twofold: the formulation of a new research topic now addressed by several teams in the world, and allowed us to start a three year collaboration with Samsung Electronics. We developed the novel concept of *mixed camera systems* combining high-resolution color cameras with low-resolution depth cameras [21], [17],[16]. Together with our start-up company 4D Views Solutions and with Samsung, we developed the first practical depth-color multiple-camera multiple-PC system and the first algorithms to reconstruct high-quality 3D content [12].

3.5. Registration, Tracking and Recognition of People and Actions

The analysis of articulated shapes has challenged standard computer vision algorithms for a long time. There are two difficulties associated with this problem, namely how to represent articulated shapes and how to devise robust registration and tracking methods. We addressed both these difficulties and we proposed a novel kinematic representation that integrates concepts from robotics and from the geometry of vision. In 2008 we proposed a method that parameterizes the occluding contours of a shape with its intrinsic kinematic parameters, such that there is a direct mapping between observed image features and joint parameters [26]. This deterministic model has been motivated by the use of 3D data gathered with multiple cameras. However, this method was not robust to various data flaws and could not achieve state-of-the-art results on standard dataset. Subsequently, we addressed the problem using probabilistic generative models. We formulated the problem of articulated-pose estimation as a maximum-likelihood with missing data and we devised several tractable algorithms [24], [23]. We proposed several expectation-maximization procedures applied to various articulated shapes: human bodies, hands, etc. In parallel, we proposed to segment and register articulated shapes represented with graphs by embedding these graphs using the spectral properties of graph Laplacians [7]. This turned out to be a very original approach that has been followed by many other researchers in computer vision and computer graphics.

SIROCCO Project-Team

3. Research Program

3.1. Introduction

The research activities on analysis, compression and communication of visual data mostly rely on tools and formalisms from the areas of statistical image modeling, of signal processing, of machine learning, of coding and information theory. Some of the proposed research axes are also based on scientific foundations of computer vision (e.g. multi-view modeling and coding). We have limited this section to some tools which are central to the proposed research axes, but the design of complete compression and communication solutions obviously rely on a large number of other results in the areas of motion analysis, transform design, entropy code design, etc which cannot be all described here.

3.2. Data Dimensionality Reduction

Manifolds, graph-based transforms, compressive sensing

Dimensionality reduction encompasses a variety of methods for low-dimensional data embedding, such as sparse and low-rank models, random low-dimensional projections in a compressive sensing framework, and sparsifying transforms including graph-based transforms. These methods are the cornerstones of many visual data processing tasks (compression, inverse problems).

Sparse representations, *compressive sensing*, and *dictionary learning* have been shown to be powerful tools for efficient processing of visual data. The objective of *sparse representations* is to find a sparse approximation of a given input data. In theory, given a dictionary matrix $A \in \mathbb{R}^{m \times n}$, and a data $\mathbf{b} \in \mathbb{R}^m$ with $m << n$ and A is of full row rank, one seeks the solution of $\min\{\|\mathbf{x}\|_0 \ : \ A\mathbf{x} = \mathbf{b}\}$, where $\|\mathbf{x}\|_0$ denotes the ℓ_0 norm of $\mathbf{x}$, i.e. the number of non-zero components in $\mathbf{x}$. A is known as the dictionary, its columns a_j are the atoms, they are assumed to be normalized in Euclidean norm. There exist many solutions x to $Ax = b$. The problem is to find the sparsest solution x, i.e. the one having the fewest nonzero components. In practice, one actually seeks an approximate and thus even sparser solution which satisfies $\min\{\|\mathbf{x}\|_0 \ : \ \|A\mathbf{x} - \mathbf{b}\|_p \leq \rho\}$, for some $\rho \geq 0$, characterizing an admissible reconstruction error.

The recent theory of *compressed sensing*, in the context of discrete signals, can be seen as an effective dimensionality reduction technique. The idea behind compressive sensing is that a signal can be accurately recovered from a small number of linear measurements, at a rate much smaller than what is commonly prescribed by the Shannon-Nyquist theorem, provided that it is sparse or compressible in a known basis. Compressed sensing has emerged as a powerful framework for signal acquisition and sensor design, with a number of open issues such as learning the basis in which the signal is sparse, with the help of dictionary learning methods, or the design and optimization of the sensing matrix. The problem is in particular investigated in the context of light fields acquisition, aiming at novel camera design with the goal of offering a good trade-off between spatial and angular resolution.

While most image and video processing methods have been developed for cartesian sampling grids, new imaging modalities (e.g. point clouds, light fields) call for representations on irregular supports that can be well represented by *graphs*. Reducing the dimensionality of such signals require designing novel transforms yielding compact signal representation. One example of transform is the Graph Fourier transform whose basis functions are given by the eigenvectors of the graph Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is a diagonal degree matrix whose i^{th} diagonal element is equal to the sum of the weights of all edges incident to the node i, and $\mathbf{A}$ the adjacency matrix. The eigenvectors of the Laplacian of the graph, also called Laplacian eigenbases, are analogous to the Fourier bases in the Euclidean domain and allow representing the signal residing on the graph as a linear combination of eigenfunctions akin to Fourier Analysis. This transform is particularly efficient for compacting smooth signals on the graph. The problems which therefore need to be addressed are (i) to define graph structures on which the corresponding signals are smooth for different imaging modalities and (ii) the design of transforms compacting well the signal energy with a tractable computational complexity.

3.3. Deep neural networks

Autoencoders, Neural Networks, Recurrent Neural Networks

From dictionary learning which we have investigated a lot in the past, our activity is now evolving towards deep learning techniques which we are considering for dimensionality reduction. We address the problem of unsupervised learning of transforms and prediction operators that would be optimal in terms of energy compaction, considering autoencoders and neural network architectures.

An autoencoder is a neural network with an encoder g_e, parametrized by θ, that computes a representation Y from the data X, and a decoder g_d, parametrized by ϕ, that gives a reconstruction $\widehat{X}$ of X (see Figure below). Autoencoders can be used for dimensionality reduction, compression, denoising. When it is used for compression, the representation need to be quantized, leading to a quantized representation $\widehat{Y} = Q(Y)$ (see Figure below). If an autoencoder has fully-connected layers, the architecture, and the number of parameters to be learned, depends on the image size. Hence one autoencoder has to be trained per image size, which poses problems in terms of genericity.

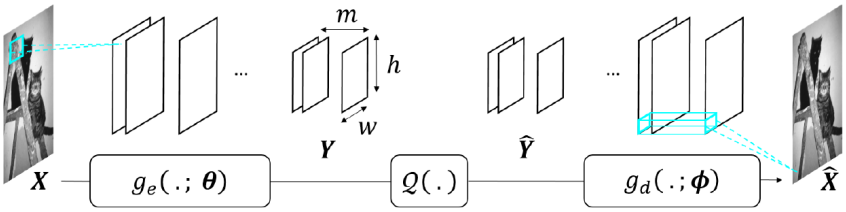


Figure 1. Illustration of an autoencoder.

To avoid this limitation, architectures without fully-connected layer and comprising instead convolutional layers and non-linear operators, forming convolutional neural networks (CNN) may be preferrable. The obtained representation is thus a set of so-called feature maps.

The other problems that we address with the help of neural networks are scene geometry and scene flow estimation, view synthesis, prediction and interpolation with various imaging modalities. The problems are posed either as supervised or unsupervised learning tasks. Our scope of investigation includes autoencoders, convolutional networks, variational autoencoders and generative adversarial networks (GAN) but also recurrent networks and in particular Long Short Term Memory (LSTM) networks. Recurrent neural networks attempting to model time or sequence dependent behaviour, by feeding back the output of a neural network layer at time t to the input of the same network layer at time t+1, have been shown to be interesting tools for temporal frame prediction. LSTMs are particular cases of recurrent networks made of cells composed of three types of neural layers called gates.

Deep neural networks have also been shown to be very promising for solving inverse problems (e.g. super-resolution, sparse recovery in a compressive sensing framework, inpainting) in image processing. Variational autoencoders, generative adversarial networks (GAN), learn, from a set of examples, the latent space or the manifold in which the images, that we search to recover, reside. The inverse problems can be re-formulated using a regularization in the latent space learned by the network. For the needs of the regularization, the learned latent space may need to verify certain properties such as preserving distances or neighborhood of the input space, or in terms of statistical modeling. GANs, trained to produce images that are plausible, are also useful tools for learning texture models, expressed via the filters of the network, that can be used for solving problems like inpainting or view synthesis.

3.4. Coding theory

OPTA limit (Optimum Performance Theoretically Attainable), Rate allocation, Rate-Distortion optimization, lossy coding, joint source-channel coding multiple description coding, channel modelization, oversampled frame expansions, error correcting codes.

Source coding and channel coding theory [0] is central to our compression and communication activities, in particular to the design of entropy codes and of error correcting codes. Another field in coding theory which has emerged in the context of sensor networks is Distributed Source Coding (DSC). It refers to the compression of correlated signals captured by different sensors which do not communicate between themselves. All the signals captured are compressed independently and transmitted to a central base station which has the capability to decode them jointly. DSC finds its foundation in the seminal Slepian-Wolf [0] (SW) and Wyner-Ziv [0] (WZ) theorems. Let us consider two binary correlated sources X and Y. If the two coders communicate, it is well known from Shannon's theory that the minimum lossless rate for X and Y is given by the joint entropy $H(X,Y)$. Slepian and Wolf have established in 1973 that this lossless compression rate bound can be approached with a vanishing error probability for long sequences, even if the two sources are coded separately, provided that they are decoded jointly and that their correlation is known to both the encoder and the decoder.

In 1976, Wyner and Ziv considered the problem of coding of two correlated sources X and Y, with respect to a fidelity criterion. They have established the rate-distortion function $R*_{X|Y}(D)$ for the case where the side information Y is perfectly known to the decoder only. For a given target distortion D, $R*_{X|Y}(D)$ in general verifies $R_{X|Y}(D) \leq R*_{X|Y}(D) \leq R_X(D)$, where $R_{X|Y}(D)$ is the rate required to encode X if Y is available to both the encoder and the decoder, and R_X is the minimal rate for encoding X without SI. These results give achievable rate bounds, however the design of codes and practical solutions for compression and communication applications remain a widely open issue.

[0] T. M. Cover and J. A. Thomas, Elements of Information Theory, Second Edition, July 2006.

[0] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources." IEEE Transactions on Information Theory, 19(4), pp. 471-480, July 1973.

[0] A. Wyner and J. Ziv, "The rate-distortion function for source coding ith side information at the decoder." IEEE Transactions on Information Theory, pp. 1-10, January 1976.

Stars Project-Team

3. Research Program

3.1. Introduction

Stars follows three main research directions: perception for activity recognition, action recognition and semantic activity recognition. **These three research directions are organized following the workflow of activity recognition systems**: First, *the perception* and *the action recognition* directions provide new techniques to extract powerful features, whereas *the semantic activity recognition* research direction provides new paradigms to match these features with concrete video analytic and healthcare applications.

Transversely, we consider a *new research axis in machine learning*, combining a priori knowledge and learning techniques, to set up the various models of an activity recognition system. A major objective is to automate model building or model enrichment at the perception level and at the understanding level.

3.2. Perception for Activity Recognition

Participants: François Brémond, Antitza Dantcheva, Sabine Moisan, Monique Thonnat.

Activity Recognition, Scene Understanding, Machine Learning, Computer Vision, Cognitive Vision Systems, Software Engineering

3.2.1. Introduction

Our main goal in perception is to develop vision algorithms able to address the large variety of conditions characterizing real world scenes in terms of sensor conditions, hardware requirements, lighting conditions, physical objects, and application objectives. We have also several issues related to perception which combine machine learning and perception techniques: learning people appearance, parameters for system control and shape statistics.

3.2.2. Appearance Models and People Tracking

An important issue is to detect in real-time physical objects from perceptual features and predefined 3D models. It requires finding a good balance between efficient methods and precise spatio-temporal models. Many improvements and analysis need to be performed in order to tackle the large range of people detection scenarios.

Appearance models. In particular, we study the temporal variation of the features characterizing the appearance of a human. This task could be achieved by clustering potential candidates depending on their position and their reliability. This task can provide any people tracking algorithms with reliable features allowing for instance to (1) better track people or their body parts during occlusion, or to (2) model people appearance for re-identification purposes in mono and multi-camera networks, which is still an open issue. The underlying challenge of the person re-identification problem arises from significant differences in illumination, pose and camera parameters. The re-identification approaches have two aspects: (1) establishing correspondences between body parts and (2) generating signatures that are invariant to different color responses. As we have already several descriptors which are color invariant, we now focus more on aligning two people detection and on finding their corresponding body parts. Having detected body parts, the approach can handle pose variations. Further, different body parts might have different influence on finding the correct match among a whole gallery dataset. Thus, the re-identification approaches have to search for matching strategies. As the results of the re-identification are always given as the ranking list, re-identification focuses on learning to rank. "Learning to rank" is a type of machine learning problem, in which the goal is to automatically construct a ranking model from a training data.

Therefore, we work on information fusion to handle perceptual features coming from various sensors (several cameras covering a large scale area or heterogeneous sensors capturing more or less precise and rich information). New 3D RGB-D sensors are also investigated, to help in getting an accurate segmentation for specific scene conditions.

Long term tracking. For activity recognition we need robust and coherent object tracking over long periods of time (often several hours in video surveillance and several days in healthcare). To guarantee the long term coherence of tracked objects, spatio-temporal reasoning is required. Modeling and managing the uncertainty of these processes is also an open issue. In Stars we propose to add a reasoning layer to a classical Bayesian framework modeling the uncertainty of the tracked objects. This reasoning layer can take into account the a priori knowledge of the scene for outlier elimination and long-term coherency checking.

Controlling system parameters. Another research direction is to manage a library of video processing programs. We are building a perception library by selecting robust algorithms for feature extraction, by insuring they work efficiently with real time constraints and by formalizing their conditions of use within a program supervision model. In the case of video cameras, at least two problems are still open: robust image segmentation and meaningful feature extraction. For these issues, we are developing new learning techniques.

3.3. Action Recognition

Participants: François Brémond, Antitza Dantcheva, Monique Thonnat.

Machine Learning, Computer Vision, Cognitive Vision Systems

3.3.1. Introduction

Due to the recent development of high processing units, such as GPU, this is now possible to extract meaningful features directly from videos (e.g. video volume) to recognize reliably short actions. Action Recognition benefits also greatly from the huge progress made recently in Machine Learning (e.g. Deep Learning), especially for the study of human behavior. For instance, Action Recognition enables to measure objectively the behavior of humans by extracting powerful features characterizing their everyday activities, their emotion, eating habits and lifestyle, by learning models from a large number of data from a variety of sensors, to improve and optimize for example, the quality of life of people suffering from behavior disorders. However, Smart Homes and Partner Robots have been well advertised but remain laboratory prototypes, due to the poor capability of automated systems to perceive and reason about their environment. A hard problem is for an automated system to cope 24/7 with the variety and complexity of the real world. Another challenge is to extract people fine gestures and subtle facial expressions to better analyze behavior disorders, such as anxiety or apathy. Taking advantage of what is currently studied for self-driving cars or smart retails, there is a large avenue to design ambitious approaches for the healthcare domain. In particular, the advance made with Deep Learning algorithms has already enabled to recognize complex activities, such as cooking interactions with instruments, and from this analysis to differentiate healthy people from the ones suffering from dementia.

To address these issues, we propose to tackle several challenges:

3.3.2. Action recognition in the wild

The current Deep Learning techniques are mostly developed to work on few clipped videos, which have been recorded with students performing a limited set of predefined actions in front of a camera with high resolution. However, real life scenarios include actions performed in a spontaneous manner by older people (including people interactions with their environment or with other people), from different viewpoints, with varying framerate, partially occluded by furniture at different locations within an apartment depicted through long untrimmed videos. Therefore, a new dedicated dataset should be collected in a real-world setting to become a public benchmark video dataset and to design novel algorithms for ADL activity recognition. A special attention should be taken to anonymize the videos.

3.3.3. Attention mechanisms for action recognition

Activities of Daily Living (ADL) and video-surveillance activities are different from internet activities (e.g. Sports, Movies, YouTube), as they may have very similar context (e.g. same background kitchen) with high intra-variation (different people performing the same action in different manners), but in the same time low inter-variation, similar ways to perform two different actions (e.g. eating and drinking a glass of water). Consequently, fine-grained actions are badly recognized. So, we will design novel attention mechanisms for action recognition, for the algorithm being able to focus on a discriminative part of the person conducting the action. For instance, we will study attention algorithms, which could focus on the most appropriate body parts (e.g. full body, right hand). In particular, we plan to design a soft mechanism, learning the attention weights directly on the feature map of a 3DconvNet, a powerful convolutional network, which takes as input a batch of videos.

3.3.4. Action detection for untrimmed videos

Many approaches have been proposed to solve the problem of action recognition in short clipped 2D videos, which achieved impressive results with hand-crafted and deep features. However, these approaches cannot address real life situations, where cameras provide online and continuous video streams in applications such as robotics, video surveillance, and smart-homes. Here comes the importance of action detection to help recognizing and localizing each action happening in long videos. Action detection can be defined as the ability to localize starting and ending of each human action happening in the video, in addition to recognizing each action label. There have been few action detection algorithms designed for untrimmed videos, which are based on either sliding window, temporal pooling or frame-based labeling. However, their performance is too low to address real-word datasets. A first task consists in benchmarking the already published approaches to study their limitations on novel untrimmed video datasets, recorded following real-world settings. A second task could be to propose a new mechanism to improve either 1) the temporal pooling directly from the 3DconvNet architecture using for instance Temporal Convolution Networks (TCNs) or 2) frame-based labeling with a clustering technique (e.g. using Fisher Vectors) to discover the sub-activities of interest.

3.3.5. View invariant action recognition

The performance of current approaches strongly relies on the used camera angle: enforcing that the camera angle used in testing is the same (or extremely close to) as the camera angle used in training, is necessary for the approach performs well. On the contrary, the performance drops when a different camera view-point is used. Therefore, we aim at improving the performance of action recognition algorithms by relying on 3D human pose information. For the extraction of the 3D pose information, several open-source algorithms can be used, such as openpose or videopose3D (from CMU or Facebook research, https://github.com/CMU-Perceptual-Computing-Lab/openpose. Also, other algorithms extracting 3d meshes can be used. To generate extra views, Generative Adversial Network (GAN) can be used together with the 3D human pose information to complete the training dataset from the missing view.

3.3.6. Uncertainty and action recognition

Another challenge is to combine the short-term actions recognized by powerful Deep Learning techniques with long-term activities defined by constraint-based descriptions and linked to user interest. To realize this objective, we have to compute the uncertainty (i.e. likelihood or confidence), with which the short-term actions are inferred. This research direction is linked to the next one, to Semantic Activity Recognition.

3.4. Semantic Activity Recognition

Participants: François Brémond, Sabine Moisan, Monique Thonnat.

Activity Recognition, Scene Understanding, Computer Vision

3.4.1. Introduction

Semantic activity recognition is a complex process where information is abstracted through four levels: signal (e.g. pixel, sound), perceptual features, physical objects and activities. The signal and the feature levels are characterized by strong noise, ambiguous, corrupted and missing data. The whole process of scene understanding consists in analyzing this information to bring forth pertinent insight of the scene and its dynamics while handling the low level noise. Moreover, to obtain a semantic abstraction, building activity models is a crucial point. A still open issue consists in determining whether these models should be given a priori or learned. Another challenge consists in organizing this knowledge in order to capitalize experience, share it with others and update it along with experimentation. To face this challenge, tools in knowledge engineering such as machine learning or ontology are needed.

Thus we work along the following research axes: high level understanding (to recognize the activities of physical objects based on high level activity models), learning (how to learn the models needed for activity recognition) and activity recognition and discrete event systems.

3.4.2. High Level Understanding

A challenging research axis is to recognize subjective activities of physical objects (i.e. human beings, animals, vehicles) based on a priori models and objective perceptual measures (e.g. robust and coherent object tracks).

To reach this goal, we have defined original activity recognition algorithms and activity models. Activity recognition algorithms include the computation of spatio-temporal relationships between physical objects. All the possible relationships may correspond to activities of interest and all have to be explored in an efficient way. The variety of these activities, generally called video events, is huge and depends on their spatial and temporal granularity, on the number of physical objects involved in the events, and on the event complexity (number of components constituting the event).

Concerning the modeling of activities, we are working towards two directions: the uncertainty management for representing probability distributions and knowledge acquisition facilities based on ontological engineering techniques. For the first direction, we are investigating classical statistical techniques and logical approaches. For the second direction, we built a language for video event modeling and a visual concept ontology (including color, texture and spatial concepts) to be extended with temporal concepts (motion, trajectories, events ...) and other perceptual concepts (physiological sensor concepts ...).

3.4.3. Learning for Activity Recognition

Given the difficulty of building an activity recognition system with a priori knowledge for a new application, we study how machine learning techniques can automate building or completing models at the perception level and at the understanding level.

At the understanding level, we are learning primitive event detectors. This can be done for example by learning visual concept detectors using SVMs (Support Vector Machines) with perceptual feature samples. An open question is how far can we go in weakly supervised learning for each type of perceptual concept (i.e. leveraging the human annotation task). A second direction is to learn typical composite event models for frequent activities using trajectory clustering or data mining techniques. We name composite event a particular combination of several primitive events.

3.4.4. Activity Recognition and Discrete Event Systems

The previous research axes are unavoidable to cope with the semantic interpretations. However they tend to let aside the pure event driven aspects of scenario recognition. These aspects have been studied for a long time at a theoretical level and led to methods and tools that may bring extra value to activity recognition, the most important being the possibility of formal analysis, verification and validation.

We have thus started to specify a formal model to define, analyze, simulate, and prove scenarios. This model deals with both absolute time (to be realistic and efficient in the analysis phase) and logical time (to benefit from well-known mathematical models providing re-usability, easy extension, and verification). Our purpose is to offer a generic tool to express and recognize activities associated with a concrete language to specify activities in the form of a set of scenarios with temporal constraints. The theoretical foundations and the tools being shared with Software Engineering aspects.

The results of the research performed in perception and semantic activity recognition (first and second research directions) produce new techniques for scene understanding and contribute to specify the needs for new software architectures (third research direction).

THOTH Project-Team

3. Research Program

3.1. Designing and learning structured models

The task of understanding image and video content has been interpreted in several ways over the past few decades, namely image classification, detecting objects in a scene, recognizing objects and their spatial extents in an image, estimating human poses, recovering scene geometry, recognizing activities performed by humans. However, addressing all these problems individually provides us with a partial understanding of the scene at best, leaving much of the visual data unexplained.

One of the main goals of this research axis is to go beyond the initial attempts that consider only a subset of tasks jointly, by developing novel models for a more complete understanding of scenes to address all the component tasks. We propose to incorporate the structure in image and video data explicitly into the models. In other words, our models aim to satisfy the complex sets of constraints that exist in natural images and videos. Examples of such constraints include: (i) relations between objects, like signs for shops indicate the presence of buildings, people on a road are usually walking or standing, (ii) higher-level semantic relations involving the type of scene, geographic location, and the plausible actions as a global constraint, e.g., an image taken at a swimming pool is unlikely to contain cars, (iii) relating objects occluded in some of the video frames to content in other frames, where they are more clearly visible as the camera or the object itself move, with the use of long-term trajectories and video object proposals.

This research axis will focus on three topics. The first is developing deep features for video. This involves designing rich features available in the form of long-range temporal interactions among pixels in a video sequence to learn a representation that is truly spatio-temporal in nature. The focus of the second topic is the challenging problem of modeling human activities in video, starting from human activity descriptors to building intermediate spatio-temporal representations of videos, and then learning the interactions among humans, objects and scenes temporally. The last topic is aimed at learning models that capture the relationships among several objects and regions in a single image scene, and additionally, among scenes in the case of an image collection or a video. The main scientific challenges in this topic stem from learning the structure of the probabilistic graphical model as well as the parameters of the cost functions quantifying the relationships among its entities. In the following we will present work related to all these three topics and then elaborate on our research directions.

- **Deep features for vision.** Deep learning models provide a rich representation of complex objects but in return have a large number of parameters. Thus, to work well on difficult tasks, a large amount of data is required. In this context, video presents several advantages: objects are observed from a large range of viewpoints, motion information allows the extraction of moving objects and parts, and objects can be differentiated by their motion patterns. We initially plan to develop deep features for videos that incorporate temporal information at multiple scales. We then plan to further exploit the rich content in video by incorporating additional cues, such as the detection of people and their body-joint locations in video, minimal prior knowledge of the object of interest, with the goal of learning a representation that is more appropriate for video understanding. In other words, a representation that is learned from video data and targeted at specific applications. For the application of recognizing human activities, this involves learning deep features for humans and their body-parts with all their spatiotemporal variations, either directly from raw video data or “pre-processed” videos containing human detections. For the application of object tracking, this task amounts to learning object-specific deep representations, further exploiting the limited annotation provided to identify the object.
- **Modeling human activities in videos.** Humans and their activities are not only one of the most frequent and interesting subjects in videos but also one of the hardest to analyze owing to the

complexity of the human form, clothing and movements. As part of this task, the Thoth project-team plans to build on state-of-the-art approaches for spatio-temporal representation of videos. This will involve using the dominant motion in the scene as well as the local motion of individual parts undergoing a rigid motion. Such motion information also helps in reasoning occlusion relationships among people and objects, and the state of the object. This novel spatio-temporal representation ultimately provides the equivalent of object proposals for videos, and is an important component for learning algorithms using minimal supervision. To take this representation even further, we aim to integrate the proposals and the occlusion relationships with methods for estimating human pose in videos, thus leveraging the interplay among body-joint locations, objects in the scene, and the activity being performed. For example, the locations of shoulder, elbow and wrist of a person drinking coffee are constrained to move in a certain way, which is completely different from the movement observed when a person is typing. In essence, this step will model human activities by dynamics in terms of both low-level movements of body-joint locations and global high-level motion in the scene.

- **Structured models.** The interactions among various elements in a scene, such as, the objects and regions in it, the motion of object parts or entire objects themselves, form a key element for understanding image or video content. These rich cues define the structure of visual data and how it evolves spatio-temporally. We plan to develop a novel graphical model to exploit this structure. The main components in this graphical model are spatio-temporal regions (in the case of video or simply image regions), which can represent object parts or entire objects themselves, and the interactions among several entities. The dependencies among the scene entities are defined with a higher order or a global cost function. A higher order constraint is a generalization of the pairwise interaction term, and is a cost function involving more than two components in the scene, e.g., several regions, whereas a global constraint imposes a cost term over the entire image or video, e.g., a prior on the number of people expected in the scene. The constraints we plan to include generalize several existing methods, which are limited to pairwise interactions or a small restrictive set of higher-order costs. In addition to learning the parameters of these novel functions, we will focus on learning the structure of the graph itself—a challenging problem that is seldom addressed in current approaches. This provides an elegant way to go beyond state-of-the-art deep learning methods, which are limited to learning the high-level interaction among parts of an object, by learning the relationships among objects.

3.2. Learning of visual models from minimal supervision

Today's approaches to visual recognition learn models for a limited and fixed set of visual categories with fully supervised classification techniques. This paradigm has been adopted in the early 2000's, and within it enormous progress has been made over the last decade.

The scale and diversity in today's large and growing image and video collections (such as, e.g., broadcast archives, and personal image/video collections) call for a departure from the current paradigm. This is the case because to answer queries about such data, it is unfeasible to learn the models of visual content by manually and precisely annotating every relevant concept, object, scene, or action category in a representative sample of everyday conditions. For one, it will be difficult, or even impossible to decide a-priori what are the relevant categories and the proper granularity level. Moreover, the cost of such annotations would be prohibitive in most application scenarios. One of the main goals of the Thoth project-team is to develop a new framework for learning visual recognition models by actively exploring large digital image and video sources (off-line archives as well as growing on-line content), and exploiting the weak supervisory signal provided by the accompanying metadata (such as captions, keywords, tags, subtitles, or scripts) and audio signal (from which we can for example extract speech transcripts, or exploit speaker recognition models).

Textual metadata has traditionally been used to index and search for visual content. The information in metadata is, however, typically sparse (e.g., the location and overall topic of newscasts in a video archive [0])

[0]For example at the Dutch national broadcast archive Netherlands Institute of Sound and Vision, with whom we collaborated in the EU FP7 project AXES, typically one or two sentences are used in the metadata to describe a one hour long TV program.

and noisy (e.g., a movie script may tell us that two persons kiss in some scene, but not when, and the kiss may occur off screen or not have survived the final cut). For this reason, metadata search should be complemented by visual content based search, where visual recognition models are used to localize content of interest that is not mentioned in the metadata, to increase the usability and value of image/video archives. *The key insight that we build on in this research axis is that while the metadata for a single image or video is too sparse and noisy to rely on for search, the metadata associated with large video and image databases collectively provide an extremely versatile source of information to learn visual recognition models.* This form of "embedded annotation" is rich, diverse and abundantly available. Mining these correspondences from the web, TV and film archives, and online consumer generated content sites such as Flickr, Facebook, or YouTube, guarantees that the learned models are representative for many different situations, unlike models learned from manually collected fully supervised training data sets which are often biased.

The approach we propose to address the limitations of the fully supervised learning paradigm aligns with "Big Data" approaches developed in other areas: we rely on the orders-of-magnitude-larger training sets that have recently become available with metadata to compensate for less explicit forms of supervision. This will form a sustainable approach to learn visual recognition models for a much larger set of categories with little or no manual intervention. Reducing and ultimately removing the dependency on manual annotations will dramatically reduce the cost of learning visual recognition models. This in turn will allow such models to be used in many more applications, and enable new applications based on visual recognition beyond a fixed set of categories, such as natural language based querying for visual content. This is an ambitious goal, given the sheer volume and intrinsic variability of the every day visual content available on-line, and the lack of a universally accepted formalism for modeling it. Yet, the potential payoff is a breakthrough in visual object recognition and scene understanding capabilities.

This research axis is organized into the following three sub-tasks:

- **Weakly supervised learning.** For object localization we will go beyond current methods that learn one category model at a time and develop methods that learn models for different categories concurrently. This allows "explaining away" effects to be leveraged, i.e., if a certain region in an image has been identified as an instance of one category, it cannot be an instance of another category at the same time. For weakly supervised detection in video we will consider detection proposal methods. While these are effective for still images, recent approaches for the spatio-temporal domain need further improvements to be similarly effective. Furthermore, we will exploit appearance and motion information jointly over a set of videos. In the video domain we will also continue to work on learning recognition models from subtitle and script information. The basis of leveraging the script data which does not have a temporal alignment with the video, is to use matches in the narrative in the script and the subtitles (which do have a temporal alignment with the video). We will go beyond simple correspondences between names and verbs relating to self-motion, and match more complex sentences related to interaction with objects and other people. To deal with the limited amount of occurrences of such actions in a single movie, we will consider approaches that learn action models across a collection of movies.
- **Online learning of visual models.** As a larger number of visual category models is being learned, online learning methods become important, since new training data and categories will arrive over time. We will develop online learning methods that can incorporate new examples for existing category models, and learn new category models from few examples by leveraging similarity to related categories using multi-task learning methods. Here we will develop new distance-based classifiers and attribute and label embedding techniques, and explore the use of NLP techniques such as skipgram models to automatically determine between which classes transfer should occur. Moreover, NLP will be useful in the context of learning models for many categories to identify synonyms, and to determine cases of polysemy (e.g. jaguar car brand v.s. jaguar animal), and merge or refine categories accordingly. Ultimately this will result in methods that are able to learn an"encyclopedia" of visual models.
- **Visual search from unstructured textual queries.** We will build on recent approaches that learn

recognition models on-the-fly (as the query is issued) from generic image search engines such as Google Images. While it is feasible to learn models in this manner in a matter of seconds, it is challenging to use the model to retrieve relevant content in real-time from large video archives of more than a few thousand hours. To achieve this requires feature compression techniques to store visual representations in memory, and cascaded search techniques to avoid exhaustive search. This approach, however, leaves untouched the core problem of how to associate visual material with the textual query in the first place. The second approach we will explore is based on image annotation models. In particular we will go beyond image-text retrieval methods by using recurrent neural networks such as Elman networks or long short-term memory (LSTM) networks to generate natural language sentences to describe images.

3.3. Large-scale learning and optimization

We have entered an era of massive data acquisition, leading to the revival of an old scientific utopia: it should be possible to better understand the world by automatically converting data into knowledge. It is also leading to a new economic paradigm, where data is a valuable asset and a source of activity. Therefore, developing scalable technology to make sense of massive data has become a strategic issue. Computer vision has already started to adapt to these changes.

In particular, very high dimensional models such as deep networks are becoming highly popular and successful for visual recognition. This change is closely related to the advent of big data. On the one hand, these models involve a huge number of parameters and are rich enough to represent well complex objects such as natural images or text corpora. On the other hand, they are prone to overfitting (fitting too closely to training data without being able to generalize to new unseen data) despite regularization; to work well on difficult tasks, they require a large amount of labelled data that has been available only recently. Other cues may explain their success: the deep learning community has made significant engineering efforts, making it possible to learn in a day on a GPU large models that would have required weeks of computations on a traditional CPU, and it has accumulated enough empirical experience to find good hyper-parameters for its networks.

To learn the huge number of parameters of deep hierarchical models requires scalable optimization techniques and large amounts of data to prevent overfitting. This immediately raises two major challenges: how to learn without large amounts of labeled data, or with weakly supervised annotations? How to efficiently learn such huge-dimensional models? To answer the above challenges, we will concentrate on the design and theoretical justifications of deep architectures including our recently proposed deep kernel machines, with a focus on weakly supervised and unsupervised learning, and develop continuous and discrete optimization techniques that push the state of the art in terms of speed and scalability.

This research axis will be developed into three sub-tasks:

- **Deep kernel machines for structured data.** Deep kernel machines combine advantages of kernel methods and deep learning. Both approaches rely on high-dimensional models. Kernels implicitly operate in a space of possibly infinite dimension, whereas deep networks explicitly construct high-dimensional nonlinear data representations. Yet, these approaches are complementary: Kernels can be built with deep learning principles such as hierarchies and convolutions, and approximated by multilayer neural networks. Furthermore, kernels work with structured data and have well understood theoretical principles. Thus, a goal of the Thoth project-team is to design and optimize the training of such deep kernel machines.
- **Large-scale parallel optimization.** Deep kernel machines produce nonlinear representations of input data points. After encoding these data points, a learning task is often formulated as a *large-scale convex optimization problem*; for example, this is the case for linear support vector machines, logistic regression classifiers, or more generally many empirical risk minimization formulations. We intend to pursue recent efforts for making convex optimization techniques that are dedicated to machine learning more scalable. Most existing approaches address scalability issues either in model size (meaning that the function to minimize is defined on a domain of very high dimension), or in the amount of training data (typically, the objective is a large sum of elementary functions). There is

thus a large room for improvements for techniques that jointly take these two criteria into account.

- **Large-scale graphical models.** To represent structured data, we will also investigate graphical models and their optimization. The challenge here is two-fold: designing an adequate cost function and minimizing it. While several cost functions are possible, their utility will be largely determined by the efficiency and the effectiveness of the optimization algorithms for solving them. It is a combinatorial optimization problem involving billions of variables and is NP-hard in general, requiring us to go beyond the classical approximate inference techniques. The main challenges in minimizing cost functions stem from the large number of variables to be inferred, the inherent structure of the graph induced by the interaction terms (e.g., pairwise terms), and the high-arity terms which constrain multiple entities in a graph.

3.4. Datasets and evaluation

Standard benchmarks with associated evaluation measures are becoming increasingly important in computer vision, as they enable an objective comparison of state-of-the-art approaches. Such datasets need to be relevant for real-world application scenarios; challenging for state-of-the-art algorithms; and large enough to produce statistically significant results.

A decade ago, small datasets were used to evaluate relatively simple tasks, such as for example interest point matching and detection. Since then, the size of the datasets and the complexity of the tasks gradually evolved. An example is the Pascal Visual Object Challenge with 20 classes and approximately 10,000 images, which evaluates object classification and detection. Another example is the ImageNet challenge, including thousands of classes and millions of images. In the context of video classification, the TrecVid Multimedia Event Detection challenges, organized by NIST, evaluate activity classification on a dataset of over 200,000 video clips, representing more than 8,000 hours of video, which amounts to 11 months of continuous video.

Almost all of the existing image and video datasets are annotated by hand; it is the case for all of the above cited examples. In some cases, they present limited and unrealistic viewing conditions. For example, many images of the ImageNet dataset depict upright objects with virtually no background clutter, and they may not capture particularly relevant visual concepts: most people would not know the majority of subcategories of snakes cataloged in ImageNet. This holds true for video datasets as well, where in addition a taxonomy of action and event categories is missing.

Our effort on data collection and evaluation will focus on two directions. First, we will design and assemble video datasets, in particular for action and activity recognition. This includes defining relevant taxonomies of actions and activities. Second, we will provide data and define evaluation protocols for weakly supervised learning methods. This does not mean of course that we will forsake human supervision altogether: some amount of ground-truth labeling is necessary for experimental validation and comparison to the state of the art. Particular attention will be payed to the design of efficient annotation tools.

Not only do we plan to collect datasets, but also to provide them to the community, together with accompanying evaluation protocols and software, to enable a comparison of competing approaches for action recognition and large-scale weakly supervised learning. Furthermore, we plan to set up evaluation servers together with leaderboards, to establish an unbiased state of the art on held out test data for which the ground-truth annotations are not distributed. This is crucial to avoid tuning the parameters for a specific dataset and to guarantee a fair evaluation.

- **Action recognition.** We will develop datasets for recognizing human actions and human-object interactions (including multiple persons) with a significant number of actions. Almost all of today's action recognition datasets evaluate classification of short video clips into a number of predefined categories, in many cases a number of different sports, which are relatively easy to identify by their characteristic motion and context. However, in many real-world applications the goal is to identify and localize actions in entire videos, such as movies or surveillance videos of several hours. The actions targeted here are "real-world" and will be defined by compositions of atomic actions into higher-level activities. One essential component is the definition of relevant taxonomies of actions

and activities. We think that such a definition needs to rely on a decomposition of actions into poses, objects and scenes, as determining all possible actions without such a decomposition is not feasible. We plan to provide annotations for spatio-temporal localization of humans as well as relevant objects and scene parts for a large number of actions and videos.

- **Weakly supervised learning.** We will collect weakly labeled images and videos for training. The collection process will be semi-automatic. We will use image or video search engines such as Google Image Search, Flickr or YouTube to find visual data corresponding to the labels. Initial datasets will be obtained by manually correcting whole-image/video labels, i.e., the approach will evaluate how well the object model can be learned if the entire image or video is labeled, but the object model has to be extracted automatically. Subsequent datasets will features noisy and incorrect labels. Testing will be performed on PASCAL VOC'07 and ImageNet, but also on more realistic datasets similar to those used for training, which we develop and manually annotate for evaluation. Our dataset will include both images and videos, the categories represented will include objects, scenes as well as human activities, and the data will be presented in realistic conditions.
- **Joint learning from visual information and text.** Initially, we will use a selection from the large number of movies and TV series for which scripts are available on-line, see for example http://www.dailyscript.com and http://www.weeklyscript.com. These scripts can easily be aligned with the videos by establishing correspondences between script words and (timestamped) spoken ones obtained from the subtitles or audio track. The goal is to jointly learn from visual content and text. To measure the quality of such a joint learning, we will manually annotate some of the videos. Annotations will include the space-time locations of the actions as well as correct parsing of the sentence. While DVDs will, initially, receive most attention, we will also investigate the use of data obtained from web pages, for example images with captions, or images and videos surrounded by text. This data is by nature more noisy than scripts.

WILLOW Team

3. Research Program

3.1. 3D object and scene modeling, analysis, and retrieval

This part of our research focuses on geometric models of specific 3D objects at the local (differential) and global levels, physical and statistical models of materials and illumination patterns, and modeling and retrieval of objects and scenes in large image collections. Our past work in these areas includes research aimed at recognizing rigid 3D objects in cluttered photographs taken from arbitrary viewpoints (Rothganger *et al.*, 2006), segmenting video sequences into parts corresponding to rigid scene components before recognizing these in new video clips (Rothganger *et al.*, 2007), retrieval of particular objects and buildings from images and videos (Sivic and Zisserman, 2003) and (Philbin *et al.*, 2007), and a theoretical study of a general formalism for modeling central and non-central cameras using the formalism and terminology of classical projective geometry (Ponce, 2009 and Batog *et al.*, 2010).

We have also developed multi-view stereopsis algorithms that have proven remarkably effective at recovering intricate details and thin features of compact objects and capturing the overall structure of large-scale, cluttered scenes. We have obtained a US patent 8,331,615 [0] for the corresponding software (PMVS, https://github.com/pmoulon/CMVS-PMVS) which is available under a GPL license and used for film production by ILM and Weta as well as by Google in Google Maps. It is also the basic technology used by Iconem, a start-up founded by Y. Ubelmann, a Willow collaborator. We have also applied our multi-view-stereo approach to model archaeological sites together with developing representations and efficient retrieval techniques to enable matching historical paintings to 3D models of archaeological sites (Russel *et al.*, 2011).

Our current efforts in this area are outlined in detail in Section. 7.1 .

3.2. Category-level object and scene recognition

The objective in this core part of our research is to learn and recognize quickly and accurately thousands of visual categories, including materials, objects, scenes, and broad classes of temporal events, such as patterns of human activities in picnics, conversations, etc. The current paradigm in the vision community is to model/learn one object category (read 2D aspect) at a time. If we are to achieve our goal, we have to break away from this paradigm, and develop models that account for the tremendous variability in object and scene appearance due to texture, material, viewpoint, and illumination changes within each object category, as well as the complex and evolving relationships between scene elements during the course of normal human activities.

Our current work in this area is outlined in detail in Section 7.2 .

3.3. Image restoration, manipulation and enhancement

The goal of this part of our research is to develop models, and methods for image/video restoration, manipulation and enhancement. The ability to “intelligently" manipulate the content of images and video is just as essential as high-level content interpretation in many applications: This ranges from restoring old films or removing unwanted wires and rigs from new ones in post production, to cleaning up a shot of your daughter at her birthday party, which is lovely but noisy and blurry because the lights were out when she blew the candles, or editing out a tourist from your Roman holiday video. Going beyond the modest abilities of current “digital zoom" (bicubic interpolation in general) so you can close in on that birthday cake, “deblock" a football game on TV, or turn your favorite DVD into a blue-ray, is just as important.

[0] The patent: "Match, Expand, and Filter Technique for Multi-View Stereopsis" was issued December 11, 2012 and assigned patent number 8,331,615.

In this context, we believe there is a new convergence between computer vision, machine learning, and signal processing. For example: The idea of exploiting self-similarities in image analysis, originally introduced in computer vision for texture synthesis applications (Efros and Leung, 1999), is the basis for non-local means (Buades *et al.*, 2005), one of today's most successful approaches to image restoration. In turn, by combining a powerful sparse coding approach to non-local means (Dabov *et al.*, 2007) with modern machine learning techniques for dictionary learning (Mairal *et al.*, 2010), we have obtained denoising and demosaicking results that are the state of the art on standard benchmarks (Mairal *et al.*, 2009).

Our current work is outlined in detail in Section 7.3 .

3.4. Human activity capture and classification

From a scientific point of view, visual action understanding is a computer vision problem that until recently has received little attention outside of extremely specific contexts such as surveillance or sports. Many of the current approaches to the visual interpretation of human activities are designed for a limited range of operating conditions, such as static cameras, fixed scenes, or restricted actions. The objective of this part of our project is to attack the much more challenging problem of understanding actions and interactions in unconstrained video depicting everyday human activities such as in sitcoms, feature films, or news segments. The recent emergence of automated annotation tools for this type of video data (Everingham, Sivic, Zisserman, 2006; Laptev, Marszałek, Schmid, Rozenfeld, 2008; Duchenne, Laptev, Sivic, Bach, Ponce, 2009) means that massive amounts of labelled data for training and recognizing action models will at long last be available.

Our research agenda in this scientific domain is described below and our recent results are outlined in detail in Section 7.4 .

- **Weakly-supervised learning and annotation of human actions in video.** We aim to leverage the huge amount of video data using readily-available annotations in the form of video scripts. Scripts, however, often provide only imprecise and incomplete information about the video. We address this problem with weakly-supervised learning techniques both at the text and image levels.
- **Descriptors for video representation.** Video representation has a crucial role for recognizing human actions and other components of a visual scene. Our work in this domain aims to develop generic methods for representing video data based on realistic assumptions. In particular, we develop deep learning methods and design new trainable representations for various tasks such as human action recognition, person detection, segmentation and tracking.

3.5. Learning embodied representations

Computer vision has come a long way toward understanding images and videos in terms of scene geometry, object labels, locations and poses of people or classes of human actions. This "understanding", however, remains largely disconnected from reasoning about the physical world. For example, what will happen if removing a tablecloth from a setted table? What actions will be needed to resume an interrupted meal? We believe that a true *embodied* understanding of dynamic scenes from visual observations is the next major research challenge. We plan to address this challenge by developing new models and algorithms with an emphasis on the synergy between vision, learning, robotics and natural language understanding. If successful, this research direction will bring significant advances in high-impact applications such as autonomous driving, home robotics and personal visual assistance.

Learning embodied representations is planned to be a major research axis for the successor of the Willow team. Meanwhile we have already started work in this direction and report our first results in Section 7.5 .