RESEARCH CENTER

FIELD
Perception, Cognition, Interaction

Activity Report 2012

Section Scientific Foundations

Edition: 2013-04-24

ALPAGE Project-Team

3. Scientific Foundations

3.1. From programming languages to linguistic grammars

Participants: Éric Villemonte de La Clergerie, Benoît Sagot, Pierre Boullier.

Historically, several members of Alpage were originally specialists in the domain of modeling and parsing for programming languages, and are working for more than 15 years on the generalization and extension of the techniques involved to the domain of natural language. The shift from programming language grammars to NLP grammars seriously increases complexity and requires ways to handle the ambiguities inherent in every human language. It is well known that these ambiguities are the sources of many badly handled combinatorial explosions.

Furthermore, while most programming languages are expressed by (subclasses) of well-understood context-free grammars (CFGs), no consensual grammatical formalism has yet been accepted by the whole linguistic community for the description of human languages. On the contrary, new formalisms (or variants of older ones) appear constantly. Many of them may be classified into the three following large families:

Mildly Context-Sensitive (MCS) formalisms They manipulate possibly complex elementary structures with enough restrictions to ensure the possibility of parsing with polynomial time complexities. They include, for instance, Tree Adjoining Grammars (TAGs) and Multi-component TAGs with trees as elementary structures, Linear Indexed Grammars (LIGs). Although they are strictly more powerful than MCS formalisms, Range Concatenation Grammars (RCGs, introduced and used by Alpage members, such as Pierre Boullier and Benoît Sagot [58], [96], [103]) are also parsable in polynomial time.

Unification-based formalisms They combine a context-free backbone with logic arguments as decoration on non-terminals. Most famous representatives are Definite Clause Grammars (DCGs) where PROLOG powerful unification is used to compute and propagate these logic arguments. More recent formalisms, like Lexical Functional Grammars (LFGs) and Head-Driven Phrasal Structure Grammars (HPSGs) rely on more expressive Typed Feature Structures (TFS) or constraints.

Unification-based formalisms with an MCS backbone The two above-mentioned characteristics may be combined, for instance by adding logic arguments or constraints to non-terminals in TAGs.

An efficient way to develop large-coverage hand-crafted symbolic grammars is to use adequate tools and adequate levels of representation, and in particular Meta-Grammars, one of Alpage's areas of expertise [121], [117]. Meta-Grammars allows the linguist to focus on a modular description of the linguistic aspects of a grammar, rather than focusing on the specific aspects of a given grammatical formalism. Translation from MGs to grammatical formalisms such as TAG or LFG may be automatically handled. Graphical environments can be used to design MGs and their modularity provides a promising way for sharing the description of common linguistic phenomena across human languages.

3.2. Statistical Parsing

Contrary to symbolic approaches to parsing, in statistical parsing, the grammar is extracted from a corpus of syntactic trees : a treebank. The main advantage of the statistical approach is to encode within the same framework the parsing and disambiguating tasks. The extracted grammar rules are associated with probabilities that allow to score and rank the output parse trees of an input sentence. This obvious advantage of probabilistic context-free grammars has long been counterbalanced by two main shortcomings that resulted in poor performance for plain PCFG parsers: (i) the generalization encoded in non terminal symbols that stand for syntagmatic phrases is too coarse (so probabilistic independence between rules is too strong an assertion) and (ii) lexical items are underused. In the last decade though, effective solutions to these shortcomings have

been proposed. Symbol annotation, either manual [84] or automatic [91], [92] captures inter-dependence between CFG rules. Lexical information is integrated in frameworks such as head-driven models that allow lexical heads to percolate up the syntagmatic tree [72], or probabilistic models derived from lexicalized Tree Adjoining grammars, such as Stochastic Tree Insertion Grammars [70].

In the same period, totally different parsing architectures have been proposed, to obtain dependency-based syntactic representations. The properties of dependency structures, in which each word is related to exactly one other word, make it possible to define dependency parsing as a sequence of simple actions (such as read buffer and store word on top of a stack, attach read word as dependent of stack top word, attach read word as governor of stack top word ...) [129], [89]. Classifiers can be trained to choose the best action to perform given a partial parsing configuration. In another approach, dependency parsing is cast into the problem of finding the maximum spanning tree within the graph of all possible word-to-word dependencies, and online classification is used to weight the edges [85]. These two kinds of statistical dependency parsing allow to benefit from discriminative learning, and its ability to easily integrate various kinds of features, which is typically needed in a complex task such as parsing.

Statistical parsing is now effective, both for syntagmatic representations and dependency-based syntactic representations. Alpage has obtained state-of-the-art parsing results for French, by adapting various parser learners for French, and works on the current challenges in statistical parsing, namely (1) robustness and portability across domains and (2) the ability to incorporate exogenous data to improve parsing attachment decisions. We review below the approaches that Alpage has tested and adapted, and the techniques that we plan to investigate to answer these challenges.

In order to investigate statistical parsers for French, we have first worked how to use the French Treebank [55] and derive the best input for syntagmatic statistical parsing [74]. Benchmarking several PCFG-based learning frameworks [11] has led to state-of-the-art results for French, the best performance being obtained with the split-merge Berkeley parser (PCFG with latent annotations) [92].
In parallel to the work on dependency based representation, presented in the next paragraph, we also conducted a preliminary set of experiments on richer parsing models based on Stochastic Tree Insertion Grammars as used in [70] and which, besides their inferior performance compared to PCFG-LA based parser, raise promising results with respect to dependencies that can be extracted from derivation trees. One variation we explored, that uses a specific TIG grammar instance, a *vertical* grammar called *spinal* grammars, exhibits interesting properties wrt the grammar size typically extracted from treebanks (a few hundred unlexicalized trees, compared to 14 000 CFG rules). These models are currently being investigated in our team [113].
Pursuing our work on PCFG-LA based parsing, we investigated the automatic conversion of the treebank into dependency syntax representations [65], that are easier to use for various NLP applications such as question-answering or information extraction, and that are a better ground for further semantic analysis. This conversion can be applied on the treebank, before training a dependency-based parser, or on PCFG-LA parsed trees. This gives the possibility to evaluate and compare on the same gold data, both syntagmatic- and dependency-based statistical parsing. This also paved the way for studies on the influence of various types of lexical information. Results are sketched in section 6.4 .

3.3. Dynamic wide coverage lexical resources

Participants: Benoît Sagot, Laurence Danlos, Rosa Stern, Éric Villemonte de La Clergerie.

Grammatical formalisms and associated parsing generators are useful only when used together with linguistic resources (lexicons, grammars) so as to build operational parsers, especially when considering modern lexically oriented grammatical formalisms. Hence, linguistic resources are the topic of the following section.

However, wide coverage linguistic resources are scarce and expensive, because they are difficult to build, especially when hand-crafted. This observation motivates us to investigate methods, along to manual development techniques, to automatically or semi-automatically acquire, supplement and correct linguistic resources.

Linguistic expertise remains a very important asset to benefit efficiently from such techniques, including those described below. Moreover, linguistically oriented environments with adequate collaborative interfaces are needed to facilitate the edition, comparison, validation and maintenance of large scale linguistic resources. Just to give some idea of the complexity, a syntactic lexicon, as described below, should provide rich information for several tens of thousands of lemma and several hundreds of thousands of forms.

Successful experiments have been conduced by Alpage members with different languages for the automatic acquisition of morphological knowledge from raw corpora [102]. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information and automatic detection of errors in the lexicon [130],[10]. At the semantic level, automatic wordnet development tools have been described [95], [123], [82], [80]. All such techniques need of course to be followed by manual validation, so as to ensure high-quality results.

For French, these techniques, and others, have lead some Alpage members to develop one of the main syntactic resources for French, the Le*fff* [98],[8], developed within the Alexina framework, as well as a wordnet for French, the WOLF [7], the first freely available resource of the kind.

In the last few years, Alpage members have shown how to benefit from other more linguistically-oriented resources, such as the Lexique-Grammaire and DICOVALENCE , in order to improve the coverage and quality of the Le*fff* and the WOLF. This work is a good example of how Inria and Paris 7 members of Alpage fruitful collaborate: this collaboration between NLP computer scientists and NLP linguists have resulted in significant advances which would have not been possible otherwise.

Moreover, an increasing effort has been made towards multilingual aspects. In particular, Alexina lexicons developed in 2010 or before exist for Slovak [102], Polish [104], English, Spanish [87], [86] and Persian [108], not including freely-available lexicons adapted to the Alexina framework.

3.4. Shallow processing

Participants: Éric Villemonte de La Clergerie, Benoît Sagot, Rosa Stern.

The constitution of resources such as lexica or grammars raises the issues of the evaluation of these resources to assess their quality and coverage. For this reason, Alpage was the leader of the PASSAGE ANR project (ended in June 2010), which is the follow-up of the EASy parsing evaluation campaign held in 2004 and conducted by team LIR at LIMSI.

However, although developing parsing techniques, grammars (symbolic or probabilistic), and lexica constitute obviously the key efforts towards deep large-scale linguistic processing, these components need to be included inside a full and robust processing chain, able to handle any text from any source. The development of such linguistic chains, such as SXPipe , is not a trivial task [6]. Moreover, when used as a preliminary step before parsers, the quality of parsers' results strongly depends on the quality of such chains. In that regard, less-standard pre-processings such as word clustering have led to promising results [112].

In fact, such processing chains are mostly used as such, and not only as pre-processing tools before parsing. They aim at performing the basic tasks that produce immediately usable results for many applications, such as tokenization, sentence segmentation, spelling correction, and, most importantly, named entity detection, disambiguation and resolution.

3.5. Discourse structures

Participants: Laurence Danlos, Charlotte Roze.

Until now, the linguistic modeling and automatic processing of sentences has been the main focus of the community. However, many applications would benefit from more large-scale approaches which go beyond the level of sentences. This is not only the case for automatic translation: information extraction/retrieval, summarizing, and other applications do need to resolve anaphora, which in turn can benefit from the availability of hierarchical discourse structures induced by discourse relations (in particular through the notion of right frontier of discourse structures). Moreover, discourse structures are required to extract sequential

(chronological, logical,...) or hierarchical representations of events. It is also useful for topic extraction, which in turns can help syntactic and semantic disambiguation.

Although supra-sentential problematics received increasing attention in the last years, there is no satisfying solution to these problems. Among them, anaphora resolution and discourse structures have a far-reaching impact and are domains of expertise of Alpage members. But their formal modeling has now reached a maturity which allows to integrate them, in a near future, inside future Alpage tools, including parsing systems inherited from Atoll.

It is well known that a text is not a random sequence of sentences: sentences are linked the ones to the others by "discourse relations", which give to the text a hierarchical structure. Traditionally, it is considered that discourse relations are lexicalized by connectors (adverbial connectors like *ensuite*, conjunctions like *parce que*), or are not lexicalized. This vision is however too simple:

- first, some connectors (in particular conjunctions of subordination) introduce pure modifiers and must not be considered as bearing discourse relations,
- second, other elements than connectors can lexicalize discourse relations, in particular verbs like *précéder / to precede* or *causer / to cause*, which have facts or fact eventualities as arguments [76].

There are three main frameworks used to model discourse structures: RST, SDRT , and, more recently, D-LTAG. Inside Alpage, Laurence Danlos has introduced D-STAG (Discourse Synchronous TAGs, [77],[5]), which subsumes in an elegant way both SDRT and RST, to the extent that SDRT and RST structures can be obtained by two different partial projections of D-STAG structures. As done in D-LTAG, D-STAG extends a lexicalized TAG analysis so as to deal with the level of discourse. D-STAG has been fully formalized, and is hence possible to implement (thanks to Synchronous TAG, or even TAG parsers), provided one develops linguistic descriptions in this formalism.

METISS Project-Team

3. Scientific Foundations

3.1. Introduction

probabilistic modeling, statistical estimation, bayesian decision theory gaussian mixture modeling, Hidden Markov Model, adaptive representation, redundant system, sparse decomposition, sparsity criterion, source separation

Probabilistic approaches offer a general theoretical framework [92] which has yielded considerable progress in various fields of pattern recognition. In speech processing in particular [89], the probabilistic framework indeed provides a solid formalism which makes it possible to formulate various problems of segmentation, detection and classification. Coupled to statistical approaches, the probabilistic paradigm makes it possible to easily adapt relatively generic tools to various applicative contexts, thanks to estimation techniques for training from examples.

A particularily productive family of probabilistic models is the Hidden Markov Model, either in its general form or under some degenerated variants. The stochastic framework makes it possible to rely on well-known algorithms for the estimation of the model parameters (EM algorithms, ML criteria, MAP techniques, ...) and for the search of the best model in the sense of the exact or approximate maximum likelihood (Viterbi decoding or beam search, for example).

More recently, Bayesian networks [94] have emerged as offering a powerful framework for the modeling of musical signals (for instance, [90], [95]).

In practice, however, the use of probabilistic models must be accompanied by a number of adjustments to take into account problems occurring in real contexts of use, such as model inaccuracy, the insufficiency (or even the absence) of training data, their poor statistical coverage, etc...

Another focus of the activities of the METISS research group is dedicated to sparse representations of signals in redundant systems [93]. The use of criteria of sparsity or entropy (in place of the criterion of least squares) to force the unicity of the solution of a underdetermined system of equations makes it possible to seek an economical representation (exact or approximate) of a signal in a redundant system, which is better able to account for the diversity of structures within an audio signal.

The topic of sparse representations opens a vast field of scientific investigation : sparse decomposition, sparsity criteria, pursuit algorithms, construction of efficient redundant dictionaries, links with the non-linear approximation theory, probabilistic extensions, etc... and more recently, compressive sensing [88]. The potential applicative outcomes are numerous.

This section briefly exposes these various theoretical elements, which constitute the fundamentals of our activities.

3.2. Probabilistic approach

probability density function, gaussian model, gaussian mixture model, Hidden Markov Model, Bayesian network, maximum likelihood, maximum a posteriori, EM algorithm, inference, Viterbi algorithm, beam search, classification, hypotheses testing, acoustic parameterisation

For several decades, the probabilistic approaches have been used successfully for various tasks in pattern recognition, and more particularly in speech recognition, whether it is for the recognition of isolated words, for the retranscription of continuous speech, for speaker recognition tasks or for language identification. Probabilistic models indeed make it possible to effectively account for various factors of variability occuring in the signal, while easily lending themselves to the definition of metrics between an observation and the model of a sound class (phoneme, word, speaker, etc...).

3.2.1. Probabilistic formalism and modeling

The probabilistic approach for the representation of an (audio) class X relies on the assumption that this class can be described by a probability density function (PDF) $P(.|X)$ which associates a probability $P(Y|X)$ to any observation Y.

In the field of speech processing, the class X can represent a phoneme, a sequence of phonemes, a word from a vocabulary, or a particular speaker, a type of speaker, a language, ... Class X can also correspond to other types of sound objects, for example a family of sounds (word, music, applause), a sound event (a particular noise, a jingle), a sound segment with stationary statistics (on both sides of a rupture), etc.

In the case of audio signals, the observations Y are of an acoustical nature, for example vectors resulting from the analysis of the short-term spectrum of the signal (filter-bank coefficients, cepstrum coefficients, time-frequency principal components, etc.) or any other representation accounting for the information that is required for an efficient separation of the various audio classes considered.

In practice, the PDF P is not accessible to measurement. It is therefore necessary to resort to an approximation $\widehat{P}$ of this function, which is usually refered to as the likelihood function. This function can be expressed in the form of a parametric model.

The models most used in the field of speech and audio processing are the Gaussian Model (GM), the Gaussian Mixture Model (GMM) and the Hidden Markov Model (HMM). But recently, more general models have been considered and formalised as graphical models.

Choosing a particular family of models is based on a set of considerations ranging from the general structure of the data, some knowledge on the audio class making it possible to size the model, the speed of calculation of the likelihood function, the number of degrees of freedom of the model compared to the volume of training data available, etc.

3.2.2. Statistical estimation

The determination of the model parameters for a given class is generally based on a step of statistical estimation consisting in determining the optimal value for model parameters.

The Maximum Likelihood (ML) criterion is generally satisfactory when the number of parameters to be estimated is small w.r.t. the number of training observations. However, in many applicative contexts, other estimation criteria are necessary to guarantee more robustness of the learning process with small quantities of training data. Let us mention in particular the Maximum a Posteriori (MAP) criterion which relies on a prior probability of the model parameters expressing possible knowledge on the estimated parameter distribution for the class considered. Discriminative training is another alternative to these two criteria, definitely more complex to implement than the ML and MAP criteria.

In addition to the fact that the ML criterion is only one particular case of the MAP criterion, the MAP criterion happens to be experimentally better adapted to small volumes of training data and offers better generalization capabilities of the estimated models (this is measured for example by the improvement of the classification performance and recognition on new data). Moreover, the same scheme can be used in the framework of incremental adaptation, i.e. for the refinement of the parameters of a model using new data observed for instance, in the course of use of the recognition system.

3.2.3. Likelihood computation and state sequence decoding

During the recognition phase, it is necessary to evaluate the likelihood function of the observations for one or several models. When the complexity of the model is high, it is generally necessary to implement fast calculation algorithms to approximate the likelihood function.

In the case of HMM models, the evaluation of the likelihood requires a decoding step to find the most probable sequence of hidden states. This is done by implementing the Viterbi algorithm, a traditional tool in the field of speech recognition. However, when the acoustic models are combined with a syntagmatic model, it is necessary to call for sub-optimal strategies, such as beam search.

3.2.4. Bayesian decision

When the task to solve is the classification of an observation into one class among several closed-set possibilities, the decision usually relies on the maximum a posteriori rule.

In other contexts (for instance, in speaker verification, word-spotting or sound class detection), the problem of classification can be formulated as a binary hypotheses testing problem, consisting in deciding whether the tested observation is more likely to be pertaining to the class under test or not pertaining to it. In this case, the decision consists in acceptance or rejection, and the problem can be theoretically solved within the framework of Bayesian decision by calculating the ratio of the PDFs for the class and the non-class distributions, and comparing this ratio to a decision threshold.

In theory, the optimal threshold does not depend on the class distribution, but in practice the quantities provided by the probabilistic models are not the true PDFs, but only likelihood functions which approximate the true PDFs more or less accurately, depending on the quality of the model of the class.

The optimal threshold must be adjusted for each class by modeling the behaviour of the test on external (development) data.

3.2.5. Graphical models

In the past years, increasing interest has focused on graphical models for multi-source audio signals, such as polyphonic music signals. These models are particularly interesting, since they enable a formulation of music modelisation in a probabilistic framework.

It makes it possible to account for more or less elaborate relationship and dependencies between variables representing multiple levels of description of a music piece, together with the exploitation of various priors on the model parameters.

Following a well-established metaphore, one can say that the graphical model expresses the notion of modularity of a complex system, while probability theory provides the glue whereby the parts are combined. Such a data structure lends itself naturally to the design of efficient general-purpose algorithms.

The graphical model framework provides a way to view a number of existing models (including HMMs) as instances of a common formalism and all of them can be addressed via common machine learning tools.

A first issue when using graphical models is the one of the model design, i.e. the chosen variables for parameterizing the signal, their priors and their conditional dependency structure.

The second problem, called the inference problem, consists in estimating the activity states of the model for a given signal in the maximum a posteriori sense. A number of techniques are available to achieve this goal (sampling methods, variational methods belief propagation, ...), whose challenge is to achieve a good compromise between tractability and accuracy [94].

3.3. Sparse representations

wavelet, dictionary, adaptive decomposition, optimisation, parcimony, non-linear approximation, pursuit, greedy algorithm, computational complexity, Gabor atom, data-driven learning, principal component analysis, independant component analysis

Over the past decade, there has been an intense and interdisciplinary research activity in the investigation of sparsity and methods for sparse representations, involving researchers in signal processing, applied mathematics and theoretical computer science. This has led to the establishment of sparse representations as a key methodology for addressing engineering problems in all areas of signal and image processing, from the data acquisition to its processing, storage, transmission and interpretation, well beyond its original applications in enhancement and compression. Among the existing sparse approximation algorithms, L1-optimisation principles (Basis Pursuit, LASSO) and greedy algorithms (e.g., Matching Pursuit and its variants) have in particular been extensively studied and proved to have good decomposition performance, provided that the sparse signal model is satisfied with sufficient accuracy.

The large family of audio signals includes a wide variety of temporal and frequential structures, objects of variable durations, ranging from almost stationary regimes (for instance, the note of a violin) to short transients (like in a percussion). The spectral structure can be mainly harmonic (vowels) or noise-like (fricative consonants). More generally, the diversity of timbers results in a large variety of fine structures for the signal and its spectrum, as well as for its temporal and frequential envelope. In addition, a majority of audio signals are composite, i.e. they result from the mixture of several sources (voice and music, mixing of several tracks, useful signal and background noise). Audio signals may have undergone various types of distortion, recording conditions, media degradation, coding and transmission errors, etc.

Sparse representations provide a framework which has shown increasingly fruitful for capturing, analysing, decomposing and separating audio signals

3.3.1. Redundant systems and adaptive representations

Traditional methods for signal decomposition are generally based on the description of the signal in a given basis (i.e. a free, generative and constant representation system for the whole signal). On such a basis, the representation of the signal is unique (for example, a Fourier basis, Dirac basis, orthogonal wavelets, ...). On the contrary, an adaptive representation in a redundant system consists of finding an optimal decomposition of the signal (in the sense of a criterion to be defined) in a generating system (or dictionary) including a number of elements (much) higher than the dimension of the signal.

Let y be a monodimensional signal of length T and D a redundant dictionary composed of $N > T$ vectors g_i of dimension T.

$$y = [y(t)]_{1 \leq t \leq T} \qquad D = \{g_i\}_{1 \leq i \leq N} \quad \text{with} \quad g_i = [g_i(t)]_{1 \leq t \leq T}$$

If D is a generating system of R^T, there is an infinity of exact representations of y in the redundant system D, of the type:

$$y(t) = \sum_{1 \leq i \leq N} \alpha_i g_i(t)$$

We will denote as $\alpha = \{\alpha_i\}_{1 \leq i \leq N}$, the N coefficients of the decomposition.

The principles of the adaptive decomposition then consist in selecting, among all possible decompositions, the best one, i.e. the one which satisfies a given criterion (for example a sparsity criterion) for the signal under consideration, hence the concept of adaptive decomposition (or representation). In some cases, a maximum of T coefficients are non-zero in the optimal decomposition, and the subset of vectors of D thus selected are refered to as the basis adapted to y. This approach can be extended to approximate representations of the type:

$$y(t) = \sum_{1 \leq i \leq M} \alpha_{\phi(i)} g_{\phi(i)}(t) + e(t)$$

with $M < T$, where ϕ is an injective function of $[1, M]$ in $[1, N]$ and where $e(t)$ corresponds to the error of approximation to M terms of $y(t)$. In this case, the optimality criterion for the decomposition also integrates the error of approximation.

3.3.2. Sparsity criteria

Obtaining a single solution for the equation above requires the introduction of a constraint on the coefficients α_i. This constraint is generally expressed in the following form :

$$\alpha^* = \arg\min_{\alpha} F(\alpha)$$

Among the most commonly used functions, let us quote the various functions L_γ :

$$L_\gamma(\alpha) = \left[\sum_{1 \leq i \leq N} |\alpha_i|^\gamma \right]^{1/\gamma}$$

Let us recall that for $0 < \gamma < 1$, the function L_γ is a sum of concave functions of the coefficients α_i. Function L_0 corresponds to the number of non-zero coefficients in the decomposition.

The minimization of the quadratic norm L_2 of the coefficients α_i (which can be solved in an exact way by a linear equation) tends to spread the coefficients on the whole collection of vectors in the dictionary. On the other hand, the minimization of L_0 yields a maximally parsimonious adaptive representation, as the obtained solution comprises a minimum of non-zero terms. However the exact minimization of L_0 is an untractable NP-complete problem.

An intermediate approach consists in minimizing norm L_1, i.e. the sum of the absolute values of the coefficients of the decomposition. This can be achieved by techniques of linear programming and it can be shown that, under some (strong) assumptions the solution converges towards the same result as that corresponding to the minimization of L_0. In a majority of concrete cases, this solution has good properties of sparsity, without reaching however the level of performance of L_0.

Other criteria can be taken into account and, as long as the function F is a sum of concave functions of the coefficients α_i, the solution obtained has good properties of sparsity. In this respect, the entropy of the decomposition is a particularly interesting function, taking into account its links with the information theory.

Finally, let us note that the theory of non-linear approximation offers a framework in which links can be established between the sparsity of exact decompositions and the quality of approximate representations with M terms. This is still an open problem for unspecified redundant dictionaries.

3.3.3. Decomposition algorithms

Three families of approaches are conventionally used to obtain an (optimal or sub-optimal) decomposition of a signal in a redundant system.

The “Best Basis” approach consists in constructing the dictionary D as the union of B distinct bases and then to seek (exhaustively or not) among all these bases the one which yields the optimal decomposition (in the sense of the criterion selected). For dictionaries with tree structure (wavelet packets, local cosine), the complexity of the algorithm is quite lower than the number of bases B, but the result obtained is generally not the optimal result that would be obtained if the dictionary D was taken as a whole.

The “Basis Pursuit” approach minimizes the norm L_1 of the decomposition resorting to linear programming techniques. The approach is of larger complexity, but the solution obtained yields generally good properties of sparsity, without reaching however the optimal solution which would have been obtained by minimizing L_0.

The “Matching Pursuit” approach consists in optimizing incrementally the decomposition of the signal, by searching at each stage the element of the dictionary which has the best correlation with the signal to be decomposed, and then by subtracting from the signal the contribution of this element. This procedure is repeated on the residue thus obtained, until the number of (linearly independent) components is equal to the dimension of the signal. The coefficients α can then be reevaluated on the basis thus obtained. This greedy algorithm is sub-optimal but it has good properties for what concerns the decrease of the error and the flexibility of its implementation.

Intermediate approaches can also be considered, using hybrid algorithms which try to seek a compromise between computational complexity, quality of sparsity and simplicity of implementation.

3.3.4. Dictionary construction

The choice of the dictionary D has naturally a strong influence on the properties of the adaptive decomposition : if the dictionary contains only a few elements adapted to the structure of the signal, the results may not be very satisfactory nor exploitable.

The choice of the dictionary can rely on a priori considerations. For instance, some redundant systems may require less computation than others, to evaluate projections of the signal on the elements of the dictionary. For this reason, the Gabor atoms, wavelet packets and local cosines have interesting properties. Moreover, some general hint on the signal structure can contribute to the design of the dictionary elements : any knowledge on the distribution and the frequential variation of the energy of the signals, on the position and the typical duration of the sound objects, can help guiding the choice of the dictionary (harmonic molecules, chirplets, atoms with predetermined positions, ...).

Conversely, in other contexts, it can be desirable to build the dictionary with data-driven approaches, i.e. training examples of signals belonging to the same class (for example, the same speaker or the same musical instrument, ...). In this respect, Principal Component Analysis (PCA) offers interesting properties, but other approaches can be considered (in particular the direct optimization of the sparsity of the decomposition, or properties on the approximation error with M terms) depending on the targeted application.

In some cases, the training of the dictionary can require stochastic optimization, but one can also be interested in EM-like approaches when it is possible to formulate the redundant representation approach within a probabilistic framework.

Extension of the techniques of adaptive representation can also be envisaged by the generalization of the approach to probabilistic dictionaries, i.e. comprising vectors which are random variables rather than deterministic signals. Within this framework, the signal $y(t)$ is modeled as the linear combination of observations emitted by each element of the dictionary, which makes it possible to gather in the same model several variants of the same sound (for example various waveforms for a noise, if they are equivalent for the ear). Progress in this direction are conditioned to the definition of a realistic generative model for the elements of the dictionary and the development of effective techniques for estimating the model parameters.

3.3.5. Compressive sensing

The theoretical results around sparse representations have laid the foundations for a new research field called compressed sensing, emerging primarily in the USA. Compressed sensing investigates ways in which we can sample signals at roughly the lower information rate rather than the standard Shannon-Nyquist rate for sampled signals.

In a nutshell, the principle of Compressed Sensing is, at the acquisition step, to use as samples a number of random linear projections. Provided that the underlying phenomenon under study is sufficiently sparse, it is possible to recover it with good precision using only a few of the random samples. In a way, Compressed Sensing can be seen as a generalized sampling theory, where one is able to trade bandwidth (i.e. number of samples) with computational power. There are a number of cases where the latter is becoming much more accessible than the former; this may therefore result in a significant overall gain, in terms of cost, reliability, and/or precision.

PAROLE Project-Team

3. Scientific Foundations

3.1. Introduction

Research in speech processing gave rise to two kinds of approaches:

- research that aims at explaining how speech is produced and perceived, and that therefore includes physiological aspects (vocal tract control), physical (speech acoustics), psychoacoustics (peripheral auditory system), and cognitive aspects (building sentences),
- research aiming at modeling the observation of speech phenomena (spectral analysis, stochastic acoustic or linguistic models).

The former research topic is motivated by the high specificity of speech among other acoustical signals: the speech production system is easily accessible and measurable (at least at first approach); acoustical equations are reasonably difficult from a mathematical point of view (with simplifications that are moderately restrictive); sentences built by speakers are governed by vocabulary and grammar of the considered language. This led acousticians to develop research aiming at generating artificial speech signals of good quality, and phoneticians to develop research aiming at finding out the origin of speech sound variability and at explaining how articulators are utilized, how sounds of a language are structured and how they influence each other in continuous speech. Lastly, that led linguists to study how sentences are built. Clearly, this approach gives rise to a number of exchanges between theory and experimentation and it turns out that all these aspects of speech cannot be mastered easily at the same time.

Results available on speech production and perception do not enable using an analysis by synthesis approach for automatic speech recognition. Automatic speech recognition thus gives rise to a second approach that consists in modeling observations of speech production and perception. Efforts focused onto the design of numerical models (first simple vectors of spectral shapes and now stochastic or neural models) of word or phoneme acoustical realizations, and onto the development of statistical language models.

These two approaches are complementary; the latter borrows theoretical results on speech from the former, which, in its turn, borrows some numerical methods. Spectral analysis methods are undoubtedly the domain where exchanges are most marked. The simultaneous existence of these two approaches is one of the particularities of speech research conducted in Nancy and we intend to enhance exchanges between them. These exchanges will probably grow in number because of new applications like: **(i)** computer aided foreign language learning which requires both reliable automatic speech recognition and fine acoustic and articulatory speech analysis, **(ii)** automatic recognition of spontaneous speech which requires robustness against noise and speaker variability.

3.2. Speech Analysis and Synthesis

Our research activities focus on acoustical and perceptual cues of speech sounds, speech modifications and acoustic-to-articulatory inversion. Our main applications concern the improvement of the oral component of language learning, speech synthesis and esophageal voices.

3.2.1. Oral comprehension

We developed tools to improve speech perception and production, and made perceptual experiments to prove their efficiency in language learning. These tools are also of interest for hearing impaired people, as well as for normally hearing people in noisy environments and also for children who learn to read (children who have language disabilities without cognitive deficit or hearing impairment and "normal" children).

3.2.1.1. Computer-assisted learning of prosody

We are studying automatic detection and correction of prosodic deviations made by a learner of a foreign language. This work implies three different tasks: (a) the detection of the prosodic entities of the learner's realization (lexical accent, intonative patterns), (b) the evaluation of the deviations, by comparison with a model, and (c) their corrections, both verbal and acoustic. This last kind of feedback is directly done on the learner's realization: the deviant prosodic cues are replaced by the prosodic cues of the reference. The identification and correction tasks use speech analysis and modification tools developed in our team.

Within the framework of a new project (see 6.1.6.2), we also investigate the impact of a language intonational characteristics on the perception and production of the intonation of a foreign language.

3.2.1.2. Phonemic discrimination in language acquisition and language disabilities

We keep working on a project concerning identification of early predictors of reading, reading acquisition and language difficulties, more precisely in the field of specific developmental disabilities : dyslexia and dysphasia. A fair proportion of those children show a weakness in phonological skills, particularly in phonemic discrimination. However, the precise nature and the origin of the phonological deficits remain unspecified. In the field of dyslexia and normal acquisition of reading, our first goal was to contribute to identify early indicators of the future reading level of children.We based our work on the longitudinal study - with 85 French children - of [45], [46] which indicates that phonemic discrimination at the beginning of kindergarten is strongly linked to success and specific failure in reading acquisition. We study now the link between oral discrimination both with oral comprehension and written compréhension. Our analyses are based on the follow up of a hundred chrildren for 4 years from kindergarten to end of grade 2 (from age 4 to age 8). Publications in progress.

3.2.1.3. Esophageal voices

It is possible for laryngectomees to learn a substitution voice: the esophageal voice. This voice is far from being natural. It is characterized by a weak intensity, a background noise that bothers listening, and a low pitch frequency. A device that would convert an esophageal voice to a natural voice would be very useful for laryngectomees because it would be possible for them to communicate more easily. Such natural voice restitution techniques would ideally be implemented in a portable device.

3.2.2. Acoustic-to-articulatory inversion

Acoustic-to-articulatory inversion aims at recovering the articulatory dynamics from speech signal that may be supplemented by images of the speaker face. Potential applications concern low bit rate speech coding, automatic speech recognition, speech production disorders assessment, articulatory investigations of phonetics, talking heads and articulatory feedback for language acquisition or learning.

Works on acoustic-to-articulatory inversion widely rely on an analysis by synthesis approach that covers three essential aspects:

Solving acoustic equations. In order to solve the acoustic equations adapted to the vocal tract, one assumes that the sound wave is a plane wave in the vocal tract and that it can be unbend. There are two families of solving methods:

(i) frequency methods through the acoustical-electrical analogy,

(ii) spatio-temporal methods, through the direct solving of finite difference equations derived from Webster equations.

Measuring the vocal tract. This represents an important obstacle because there does not exist any reliable method enabling a precise measurement in time and dimension. MRI (Magnetic Resonance Imaging) enables 3D measurements but is not sufficiently fast and X-rays only allows a sagittal slice of the vocal tract to be captured while involving not acceptable health hazards.

Articulatory modeling. Articulatory models aim at describing all the possible vocal tract shapes with a small number of parameters, while preserving deformations observed on a real vocal tract. Present articulatory models often derive from data analysis of cineradiography moving pictures. One of the most widely used is the one built by Maeda [52].

One of the major difficulties of inversion is that an infinity of vocal tract shapes can give rise to the same speech spectrum. Acoustic-to-articulatory inversion methods are categorized into two families:

- methods that optimize a function generally combining speaker's articulatory effort and acoustical distance between natural and synthesized speech. They exploit constraints allowing the number of possible vocal tract shapes to be reduced.
- table look-up methods resting on an articulatory codebook of articulatory shapes indexed by their acoustical parameters (generally formant frequencies). After possible shapes have been recovered at each time, an optimization procedure is used to find an inverse solution in the form of an optimal articulatory path.

As our contribution only concerns inversion, we accepted widely used articulatory synthesis methods. We therefore chose Maeda's articulatory model, the acoustical-electrical analogy to compute the speech spectrum and the spatio-temporal method proposed by Maeda to generate the speech signal. As regards inversion, we chose Maeda's model to constrain vocal tract shapes because this model guarantees that synergy and compensation articulatory phenomena are still possible, and consequently, that articulatory deformations close to those of a human speaker may be recovered. The most important challenges in this domain are the inversion of any class of speech sounds and to perform inversion from standard spectral data, MFCC for instance. Indeed at present, only vowels and sequences of vowels can be inverted, and only some attempts concern fricatives sounds. Moreover, most of the inversion techniques use formant frequencies as input data although formants cannot be extracted from speech easily and reliably.

3.2.3. Strategies of labial coarticulation

The investigation of labial coarticulations strategies is a crucial objective with the view of developing a talking head which would be understandable by lip readers, especially deaf persons.

In the long term, our goal is to determine a method of prediction of labial coarticulation adaptable to a virtual speaker. Predicting labial coarticulation is a difficult problem that gave rise to many studies and models. To predict the anticipatory coarticulation gestures (see [42] for an overall presentation of labial coarticulation), three main models have been proposed: the look-ahead model, the time-locked model and the hybrid model.

These models were often compared on their performance in the case of the prediction of anticipation protrusion in VCV or VCCV sequences where the first vowel is unrounded, the consonant(s) is neutral with respect to labial articulation and the last vowel is rounded. There is no general agreement about the efficiency of these models. More recent models have been developed. The one of Abry and Lallouache [37] advocates for the theory of expansion movements: the movement tends to be anticipated when no phonological constraint is imposed on labiality. Cohen and Massaro [40] proposed dominance functions that require a substantial numerical training.

Most of these models derive from the observations of a limited number of speakers. We are thus developing a more explicative model, i.e., essentially a phonetically based approach that tries to understand how speakers manage to control labial parameters from the sequence of phonemes to be articulated.

3.2.4. Speech Synthesis

Data-driven speech synthesis is widely adopted to develop Text-to-Speech (TTS) synthesis systems. Basically, it consists of concatenating pieces of signal (units) selected from a pre-recorded sentence corpus. Our ongoing work on acoustic TTS was recently extended to study acoustic-visual speech synthesis (bimodal units).

3.2.4.1. Text-to-speech synthesis

Data-driven text-to-speech synthesis is usually composed of three steps to transform a text in speech signal. The first step is Natural Language Processing (NLP) which tags and analyzes the input text to obtain a set of features (phoneme sequence, word grammar categories, syllables...). It ends with a prosodic model which transforms these features into acoustic or symbolic features (F0, intensity, tones...). The second step uses a Viterbi algorithm to select units from a corpus recorded beforehand, which have the closest features to the prosodic features expected. The last step amounts to concatenate these units.

Such systems usually generate a speech signal with a high intelligibility and a naturalness far better than that achieved by old systems. However, building such a system is not an easy task [39] and the global quality mainly relies on the quality of the corpus and prosodic model. The prosodic model generally provides a good standard prosody, but, the generated speech can suffer from a lack of variability. Especially during the synthesis of extended passages, repetition of similar prosodic patterns can lead to a monotonous effect. Therefore, to avoid this problem due to the projection of linguistic features onto symbolic or acoustic dimensions (during NLP), we [41] proposed to perform the unit selection directly from linguistic features without incorporating any prosodic information. To compensate the lack of prosodic prediction, the selection needs to be performed with numerous linguistic features. The selection is no longer restrained by a prosodic model but only driven by weighted features. The consequence is that the quality of synthesis may drop in crucial instants. Our works deal to overcome this new problem while keeping advantage of the lack of prosodic model.

These works have an impact on the construction of corpus and on the NLP engine which needs to provide as much information as possible to the selection step. For instance, we introduced a chunker (shallow parser) to give us information on a potential rhythmic structure. Moreover, to perform the selection, we developed an algorithm to automatically weight the linguistic features given by the NLP. Our method relies on acoustic clustering and entropy information [41]. The originality of our approach leads us to design a more flexible unit selection step, constrained but not restrained.

3.2.4.2. Acoustic-visual speech synthesis

Audiovisual speech synthesis can be achieved using 3D features of the human face supervised by a model of speech articulation and face animation. Coarticulation is approximated by numerical models that describe the synergy of the different articulators. Acoustic signal is usually synthetic or natural speech synchronized with the animation of the face. Some of the audiovisual speech systems are inspired by recent development in speech synthesis based on samples and concatenative techniques. The main idea is to concatenate segments of recorded speech data to produce new segments. Data can be video or motion capture. The main drawback of these methods is that they focus on one field, either acoustic or visual. But (acoustic) speech is actually generated by moving articulators, which modify the speaker's face. Thus, it is natural to find out that acoustic and face movements are correlated. A key point is therefore to guarantee the internal consistency of the acoustic-visual signal so that the redundancy of these two signals acknowledged as a determining perceptive factor, can really be exploited by listeners. It is thus important to deal with the two signals (acoustic and visual) simultaneously and to keep this link during the whole process. This is why we make the distinction between audiovisual speech synthesis (where acoustic is simply synchronized with animation) and acoustic-visual speech where speech is considered as a bimodal signal (acoustic and visual) as considered in our work. Our long-term goal is to contribute to the fields of acoustic speech synthesis and audiovisual speech synthesis by building a bimodal corpus and developing an acoustic-visual speech synthesis system using bimodal unit concatenation.

3.3. Automatic speech recognition

Automatic speech recognition aims at reproducing the cognitive ability of humans to recognize and understand oral speech. Our team has been working on automatic speech recognition for decades. We began with knowledge-based recognition systems and progressively made our research works evolve towards stochastic approaches, both for acoustic and language models. Regarding acoustic models, we have especially investigated HMM (Hidden Markov Models), STM (Stochastic Trajectory Models), multi-band approach and BN (Bayesian Networks). Regarding language models, our main interest has concerned ngram approaches (word classes, trigger, impossible ngram, etc).

The main challenge of automatic speech recognition is its robustness to multiple sources of speech variability [44]. Among them, we have been focusing on acoustic environment, inter- and intra-speaker variability, different speaking styles (prepared speech, spontaneous, etc.) and non-native pronunciations.

Another specifity of automatic speech recognition is the necessity to combine efficiently all the research works (in acoustic modeling, langage modeling, speaker adaptation, etc.) into a core platform in order to evaluate them, and to go beyond pure textual transcriptions by enriching them with punctuation, syntax, etc., in order to make them exploitable by both humans and machines.

3.3.1. Acoustic features and models

The raw acoustic signal needs to be parameterized to extract the speech information it contains and to reduce its dimensionality. Most of our research and recognition technologies make use of the classical Mel Feature Cepstral Coefficients, which have proven since many years to be amongst the most efficient front-end for speech recognition. However, we have also explored alternative parameterizations to support some of our recent research progresses. For example, prosodic features such as intonation curves and vocal energy give important cues to recognize dialog acts, and more generally to compute information that relates to supra-phonemic (linguistic, dialog, ...) characteristics of speech. Prosodic features are developed jointly for both the Speech Analysis and Speech Recognition topics. We also developed a new robust front-end, which is based on wavelet-decomposition of the speech signal.

Concerning acoustic models, stochastic models are now the most popular approach for automatic speech recognition. Our research on speech recognition also largely exploits Hidden Markov Models (HMM). In fact, HMMs are mainly used to model the acoustic units to be recognized (usually triphones) in all of our recognition engines (ESPERE, ANTS...). Besides,we have investigated Bayesian Networks (BN) to explicitly represent random variables and their independence relationships to improve noise robustness.

3.3.2. Robustness and invariance

Part of our research activities about ASR aims at improving the robustness of recognizers to the different sources of variability that affect the speech signal and damage the recognition. Indeed, the issue of the lack of robustness of state-of-the-art ASR systems is certainly the most problematic one that still prevents the wide deployment of speech recognizers nowadays. In the past, we developed a large range of techniques to address this difficult topic, including robust acoustic models (such as stochastic trajectory and multi-band models) and model adaptation techniques (such as missing data theory). The following state-of-the-art approaches thus form our baseline set of technologies: MLLR (Maximum Likelihood Linear Regression), MAP (Maximum A Posteriori), PMC (Parallel Model Combination), CMN (Cepstral Mean Normalization), SAT (Speaker Adaptive Training), HLDA (Heteroscedastic Linear Discriminant Analysis), Spectral Subtraction and Jacobian Adaptation.

These technologies constitute the foundations of our recent developments in this area, such as non-native speaker adaptation, out-of-vocabulary words detection and adaptation to pronunciation variations. Handling speech variabilities may also benefit from exploiting additional external or contextual sources of information to more tightly guide the speech decoding process. This is typically the role of the language model, which shall in this context be augmented with higher-level knowledge, such as syntactic or semantic cues. Yet, automatically extracting such advanced features is very challenging, especially on imperfect transcribed speech.

The performance of automatic speech recognition (ASR) systems drastically drops when confronted with non-native speech. If we want to build an ASR system that takes into account non-native speech, we need to modify the system because, usually, ASR systems are trained on standard phone pronunciations and designed to recognize only native speech. In this way, three method categories can be applied: acoustic model transformation, pronunciation modeling and language modeling. Our contribution concerns the first two methods.

3.3.3. Segmentation

Audio indexing and automatic broadcast news transcription need the segmentation of the audio signal. The segmentation task consists in two steps: firstly, homogeneous segments are extracted and classified into speech, noise or music, secondly, speakers turns are detected in the extracted speech segments.

Speech/music segmentation requires to extract discriminant acoustic parameters. Our contribution concerns the MFCC and wavelet parameters. Another point is to find a good classifier. Various classifiers are commonly used: k-Nearest-Neighbors, Hidden Markov Models, Gaussian Mixture Models, Artificial Neural Networks.

As to detect speaker turns, the main approach consists of splitting the audio signal into segments that are assumed to contain only one speaker and then a hierarchical clustering scheme is performed for merging segments belonging to the same speaker.

3.3.4. Speech/text alignment

Speech/text alignment consists in finding time boundaries of words or phones in the audio signal knowing the orthographic transcription. The main applications of speech/text alignment are training of acoustic models, segmentation of audio corpus for building units for speech synthesis or segmentation of the sentence uttered by a learner of a foreign language. Moreover, speech/text alignement is a useful tool for linguistic researchers.

Speech/text alignment requires two steps. The first step generates the potential pronunciations of the sentence dealing with multiple pronunciations of proper nouns, liaisons, phone deletions, and assimilations. For that, the phonetizer is based on a phonetic lexicon, and either phonological rules or an automatic classifier as a decision tree. The second step finds the best pronunciation corresponding to the audio signal using acoustic HMM models and an alignment algorithm. The speech team has been working on this domain for a long time.

3.4. Speech to Speech Translation and Langage Modeling

Speech-to-Speech Translation aims at translating a source speech signal into a target speech signal. A sequential way to adress this problem is to first translate a text to another one. And after, we can connect a speech recognition system at the input and a text to speech synthesis system at the output. Several ways to adress this issue exist. The concept used in our group is to let the computer learning from a parallel text all the associations between source and target units. A unit could be a word or a phrase. In the early 1990s [38] proposes five statistical translation models which became inescapable in our community. The basic idea of the model 1 is to consider that any word of the target language could be a potential translation of any source word. The problem is then to estimate the distribution probability of a target word given a source one. The translation problem is similar to the speech recognition one. Indeed, we have to seek the best foreign sentence given a source one. This one is obtained by decoding a lattice translation in which a language and translation models are used. Several issues have to be supported in machine translation as described below.

3.4.1. Word translation

The first translation systems identify one-to-one associations between words of target and source languages. This is still necessary in the present machine translation systems. In our group we develop a new concept to learn the translation table. This approach is based on computing all the inter-lingual triggers inside a parallel corpus. This leads to a pertinent translation table [51]. Obviously, this is not sufficient in order to make a realistic translation because, with this approach, one word is always translated into one word. In fact, it is possible to express the same idea in two languages by using different numbers of words. Thus, a more general one-to-one alignement has to be achieved.

3.4.2. Phrase translation

The human translation is a very complex process which is not only word-based. A number of research groups developed phrase-based systems which are different from the baseline IBM's model in training. These methods, deals with linguistic units which consists in more than one word. The model supporting phrase-based machine translation uses reordering concept and additional feature functions. In order to retrieve phrases, several approaches have been proposed in the litterature. Most of them require word-based alignments. For example, Och and al. [54] collected all phrase pairs that were consistent with the word alignment provided by Brown's models.

We developed a phrase based algorithm which is based on finding first an adequate list of phrases. Then, we find out the best corresponding translations by using our concept of inter-lingual triggers. A list of the best translations of sequences is then selected by using simulated annealing algorithm.

3.4.3. Language model

A language model has an important role in a statistical machine translation. It ensures that the translated words constitute a valid linguistic sentence. Most of the community uses n-grams models, that is what we do also.

3.4.4. Decoding

The translation issue is treated as an optimization problem. Translating a sentence from English into a Foreign language involves finding the best Foreign target sentence f^* which maximizes the probability of f given the English source sentence e. The Bayes rule allows to formulate the probability $P(f|e)$ as follows:

$$f^* = \arg\max_f P(f|e) = \arg\max_f P(e|f)P(f)$$

The international community uses either PHARAOH [48] or MOSES [47] based on a beam search algorithm. In our group we started decoding by PHARAOH but we moved recently to MOSES.

SEMAGRAMME Team

3. Scientific Foundations

3.1. Fondation

The present proposal relies on deep mathematical foundations. We intend 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.

3.1.1. 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.

3.1.2. 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 capture 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.

3.1.3. 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, [28] 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).

ALICE Project-Team

3. Scientific Foundations

3.1. Introduction

Computer Graphics is a quickly evolving domain of research. These last few years, both acquisition techniques (e.g., range laser scanners) and computer graphics hardware (the so-called GPU's, for Graphics Processing Units) have made considerable advances. However, despite these advances, fundamental problems still remain open. For instance, a scanned mesh composed of hundred million triangles cannot be used directly in real-time visualization or complex numerical simulation. To design efficient solutions for these difficult problems, ALICE studies two fundamental issues in Computer Graphics:

- the representation of the objects, i.e., their geometry and physical properties;
- the interaction between these objects and light.

Historically, these two issues have been studied by independent research communities. However, we think that they share a common theoretical basis. For instance, multi-resolution and wavelets were mathematical tools used by both communities [25]. We develop a new approach, which consists in studying the geometry and lighting from the *numerical analysis* point of view. In our approach, geometry processing and light simulation are systematically restated as a (possibly non-linear and/or constrained) functional optimization problem. This type of formulation leads to algorithms that are more efficient. Our long-term research goal is to find a formulation that permits a unified treatment of geometry and illumination over this geometry.

3.2. Geometry Processing for engineering

Participants: Laurent Alonso, Dobrina Boltcheva, Alejandro Galindo, Phuong Ho, Samuel Hornus, Thomas Jost, Bruno Lévy, David Lopez, Romain Merland, Vincent Nivoliers, Jeanne Pellerin, Nicolas Ray, Dmitry Sokolov, Rhaleb Zayer.

Mesh processing, parameterization, splines

Geometry processing recently emerged (in the middle of the 90's) as a promising strategy to solve the geometric modeling problems encountered when manipulating meshes composed of hundred millions of elements. Since a mesh may be considered to be a *sampling* of a surface - in other words a *signal* - the *digital signal processing* formalism was a natural theoretic background for this subdomain (see e.g., [26]). Researchers of this domain then studied different aspects of this formalism applied to geometric modeling.

Although many advances have been made in the geometry processing area, important problems still remain open. Even if shape acquisition and filtering is much easier than 30 years ago, a scanned mesh composed of hundred million triangles cannot be used directly in real-time visualization or complex numerical simulation. For this reason, automatic methods to convert those large meshes into higher level representations are necessary. However, these automatic methods do not exist yet. For instance, the pioneer Henri Gouraud often mentions in his talks that the *data acquisition* problem is still open. Malcolm Sabin, another pioneer of the "Computer Aided Geometric Design" and "Subdivision" approaches, mentioned during several conferences of the domain that constructing the optimum control-mesh of a subdivision surface so as to approximate a given surface is still an open problem. More generally, converting a mesh model into a higher level representation, consisting of a set of equations, is a difficult problem for which no satisfying solutions have been proposed. This is one of the long-term goals of international initiatives, such as the AIMShape European network of excellence.

Motivated by gridding application for finite elements modeling for oil and gas exploration, in the frame of the Gocad project, we started studying geometry processing in the late 90's and contributed to this area at the early stages of its development. We developed the LSCM method (Least Squares Conformal Maps) in cooperation with Alias Wavefront [5]. This method has become the de-facto standard in automatic unwrapping, and was adopted by several 3D modeling packages (including Maya and Blender). We experimented various applications of the method, including normal mapping, mesh completion and light simulation [2].

However, classical mesh parameterization requires to partition the considered object into a set of topological disks. For this reason, we designed a new method (Periodic Global Parameterization) that generates a continuous set of coordinates over the object [6]. We also showed the applicability of this method, by proposing the first algorithm that converts a scanned mesh into a Spline surface automatically [4].

We are still not fully satisfied with these results, since the method remains quite complicated. We think that a deeper understanding of the underlying theory is likely to lead to both efficient and simple methods. For this reason, we studied last year several ways of discretizing partial differential equations on meshes, including Finite Element Modeling and Discrete Exterior Calculus. This year, we also explored Spectral Geometry Processing and Sampling Theory (more on this below).

3.3. Computer Graphics

Participants: Sylvain Lefebvre, Samuel Hornus, Bruno Lévy, Vincent Nivoliers, Nicolas Ray, Dmitry Sokolov, Rhaleb Zayer.

texture synthesis, texture mapping,

Content creation is one of the major challenge in Computer Graphics. Modelling geometries and surface appearances which are visually appealing and at the same time enforce precise design constraints is a task only accessible to highly skilled and trained designers.

In this context the team focuses on methods for by-example content creation. Given an input example and a set of constraints, we design algorithms that can automatically generate a new shape (geometry+texture). We formulate the problem of content synthesis as the joint optimization of several objectives: Preserving the local appearance of the example, enforcing global objectives (size, symmetries, mechanical properties), reaching user defined constraints (locally specified geometry, contacts). This results in a wide range of optimization problems, from statistical approaches (Markov Random fields), to combinatorial and linear optimization techniques.

In addition to the core algorithm we also work on the representation of the content, so as to allow for its efficient manipulation. In this context we develop data-structures and algorithms targeted at massively parallel architectures, such as GPUs. These are critical to reach the interactive rates expected from a content creation technique. We also propose novel ways to store and access content stored along surfaces [7] or in volumes [1].

AVIZ Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

The scientific foundations of Visual Analytics lie primarily in the domains of Information 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) [62], graphic designers such as Bertin [51] and Tufte [61], and HCI researchers in the field of Information Visualization [50].

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 [56] and Triesman's "preattentive processing" theory [60]. 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 [54]. 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 [52]. 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 [8] [59], [57], [58], [55]. 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.

IMAGINE Team

3. Scientific Foundations

3.1. A failure of standard modeling techniques?

Surprisingly, in our digital age, conceptual design of static shapes, motion and stories is almost never done on computers. Designers prefer to use traditional media even when a digital model is eventually created for setups such as industrial prototyping, and even when the elements to be designed are aimed at remaining purely virtual, such as in 3D films or games. In his keynote talk at SIGGRAPH Asia 2008, Rob Cook, vice president of technology at Pixar Animation Studios, stressed that even trained computer artists tend to avoid the use of 3D computerized tools whenever possible. They use first pen and paper, and then clay to design shapes; paper to script motion; and hand-sketched storyboards to structure narrative content and synchronise it with speech and music. Even lighting and dramatic styles are designed using 2D painting tools. The use of 3D graphics is avoided as much as possible at all of these stages, as if one could only reproduce already designed material with 3D modelling software, but not create directly with it. This disconnect can be thought of as the number one failure of digital 3D modelling methodologies. As Cook stressed: "*The new grand challenge in Computer Graphics is to make tools as transparent to the artists as special effects were made transparent to the general public*" (Cook 2008). The failure does not only affect computer artists but many users, from engineers and scientists willing to validate their ideas on virtual prototypes, to media, educators and the general public looking for simple tools to quickly personalize their favourite virtual environment.

Analyzing the reasons for this failure we observe that 3D modeling methodologies did not evolve much in the last 20 years. Standard software, such as Maya and 3dsMax, provide sophisticated interfaces to fully control all degrees of freedom and bind together an increasing number of shape and motion models. Mastering this software requires years of training to become skilled. Users have to choose the best suited representation for each individual element they need to create, and fully design a shape before being able to define its motion. In many cases, neither descriptive models, which lack high level constraints and leave the quality of results in user's hands, nor procedural ones, where realistic simulation comes at the price of control, are really convenient. A good example is modelling of garments for virtual characters. The designer may either sculpt the garment surface at rest, which provides direct control on the folds but requires lots of skill due to the lack of constraints (such as enforcing a cloth surface to be developable onto a plane), or they can tune the parameters of a physically-based model simulating cloth under gravity, which behaves as a black box and may never achieve the expected result. No mechanism is provided to roughly draft a shape, and help the user progressively improve and refine it.

Capture and reconstruction of real-world objects, using either 3D scanners or image-based methods, provides an appealing alternative for quickly creating 3D models and attracted a lot of attention from both Computer Graphics and Computer Vision research communities the last few years. Similarly, techniques for capture and reuse of real motion, enabling an easy generation of believable animation content, were widely investigated. These efforts are much welcome, since being able to embed existing objects and motion in virtual environments is extremely useful. However, it is not sufficient. One cannot scan every blade of grass, or even every expressive motion, to create a convincing virtual world. What if the content to be modelled does not exist yet, or will never exist? One of the key motivations for using digital modelling in the first place is as a tool for bringing to life new, imaginary content.

3.2. Long term vision: an "expressive virtual pen" for animated 3D content

Stepping back and taking a broader viewpoint, we observe that humans need a specialized medium or tool, such as pen and paper or a piece of clay, to convey shapes, and more generally animated scenes. Pen and paper, probably the most effective media to use, requires sketching from different viewpoints to fully represent a shape and requires a large set of drawings over time to communicate motion and stories.

Could digital modeling be turned into a tool, even more expressive and simpler to use than a pen, to quickly convey and refine shapes, motions, and stories?

This is the long term vision towards which we would like to advance.

3.3. Methodology: "Control to the user, Knowledge to the system"

Thinking of future digital modeling technologies as an "expressive virtual pen", enabling to seamlessly design, refine and convey animated 3D content, is a good source of inspiration. It led us to the following methodology:

- As when they use a pen, users should not be restricted to the editing of preset shapes or motion, but should get a **full control over their design**. This control should ideally be as easy and intuitive as when sketching, which leads to the use of gestures – although not necessarily sketching gestures – rather than of standard interfaces with menus, buttons and sliders. Ideally, these control gestures should drive the choice of the underlying geometric model, deformation tool, and animation method in a predictable but transparent way, enabling users to concentrate on their design.
- Secondly, similarly to when they draw in real, users should only have to **suggest** the 3D nature of a shape, the presence of repetitive details, or the motion or deformations that are taking place: this will allow for faster input and enable coarse to fine design, with immediate visual feedback at every stage. The modeling system should thus act similarly to a human viewer, who can imagine a 3D shape in motion from very light input such as a raw sketch. Therefore, as much as possible **a priori knowledge** should be incorporated into the models and used for inferring the missing data, leading to the use of high-level representations enabling procedural generation of content. Note that such models will also help the user towards high-quality content, since they will be able to maintain specific geometric or physical laws. Since this semi-automatic content generation should not spoil user's creativity and control, editing and refinement of the result should be allowed throughout the process.
- Lastly, creative design is indeed a matter of trial and error. We believe that creation more easily takes place when users can immediately see and play with a first version of what they have in mind, serving as support for refining their thoughts. Therefore, important features towards effective creation are to provide **real-time response** at every stage, as well as to help the user exploring the content they have created thanks to intelligent cameras and other cinematography tools.

To advance in these directions, we believe that models for shape, motion and cinematography need to be re-thought from a user centered perspective. We borrowed this concept from the Human Computer Interaction domain, but we are not referring here to **user-centred system design** (Norman 86). We rather propose to extend the concept, and develop user-centred graphical models: Ideally, a user-centred model should be designed to behave, under editing actions, the way a human user would have predicted. Editing actions may be for instance 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 used to transfer of some features from existing models to other ones. User-centred models need to incorporate knowledge in order to seamlessly generate the appropriate content from such actions. Knowledge may be for instance about developability to model paper or cloth; about constant volume to deform virtual clay or animate plausible organic shapes; about physical laws to control passive objects; or about film editing rules to generate semi-autonomous camera with planning abilities.

These user-centred models will be applied to the development of various interactive creative systems, not only for static shapes, but also for motion and stories. Although unusual, we believe that thinking about these different types of content in a similar way will enable us to improve our design principles thanks to cross fertilization between domains, and allow for more thorough experimentation and validation. The expertise we developed in our previous research team EVASION, namely the combination of layered models, adaptive degrees of freedom, and GPU computations for interactive modeling and animation, will be instrumental to ensure real-time performances. Rather than trying to create a general system that would solve everything, we plan to develop specific applications (serving as case studies), either brought by the available expertise in our

research group or by external partners. This way, user expectations should be clearly defined and final users will be available for validation. Whatever the application, we expect the use of knowledge-based, user-centred models driven by intuitive control gesture to increase both the efficiency of content creation and the quality of results.

3.4. Validation methodology

When developing digital creation tools, validation is a major challenge. Researchers working on ground-truth reconstruction can apply standard methodologies to validate their techniques, such as starting by testing the method on a representative series of toy models, for which the model to reconstruct is already known. In contrast, it is not obvious how to prove that a given tool for content creation brings a new contribution. Our strategy to tackle the problem is threefold:

- Most of our contributions will address the design of new models and algorithms for geometry and animation. Validating them will be done, as usual in Computer Graphics, by showing for instance that our method solves a problem never solved before, that the model is more general, or the computations more efficient, than using previous methods.
- Interaction for interactive content creation & editing will rely as much as possible on preliminary user studies telling us about user expectations, and on interaction paradigms and design principles already identified and validated by the HCI community. When necessary, we intend to develop as well new interaction paradigms and devices (such as the hand-navigator we are currently experimenting) and validate them through user studies. All this interaction design work will be done in collaboration with the HCI community. We already set up a long term partnership with the IIHM group from LIG in Grenoble, through the INTUACTIVE project at Grenoble INP (2011-2014) which involves co-advised students, and through the co-direction of the action "Authoring Augmented Reality" of the larger Labex PERSYVAL project (2012 – 2020).
- Lastly, working on specific applications in the domains we listed in Section 3 is essential for validation since it will give us some test beds for real-size applications. The expert users involved will be able to validate the use of our new design framework compared to their usual pipeline, both in terms of increased efficiency, and of satisfaction with new functionalities and final result. In addition to our work with scientific and industrial partners, we are establishing collaborations with the Ecole Nationale Supérieure des Arts Décoratifs (ENSAD Paris, Prof Pierre Hénon) and with the Ecole Nationale Supérieure Louis Lumière (Prof. Pascal Martin) for the evaluation of our ongoing work in shape and motion design, and on virtual cinematography.

IN-SITU Project-Team

3. Scientific Foundations

3.1. Multi-disciplinary Research

INSITU uses a multi-disciplinary research approach, including computer scientists, psychologists and designers. Working together requires an understanding of each other's methods. Much of computer science relies on formal theory, which, like mathematics, is evaluated with respect to its internal consistency. The social sciences are based more on descriptive theory, attempting to explain observed behaviour, without necessarily being able to predict it. The natural sciences seek predictive theory, using quantitative laws and models to not only explain, but also to anticipate and control naturally occurring phenomena. Finally, design is based on a corpus of accumulated knowledge, which is captured in design practice rather than scientific facts but is nevertheless very effective.

Combining these approaches is a major challenge. We are exploring an integrative approach that we call *generative theory*, which builds upon existing knowledge in order to create new categories of artefacts and explore their characteristics Our goal is to produce prototypes, research methods and software tools that facilitate the design, development and evaluation of interactive systems [40].

MANAO Team

3. Scientific Foundations

3.1. Related Scientific Domains

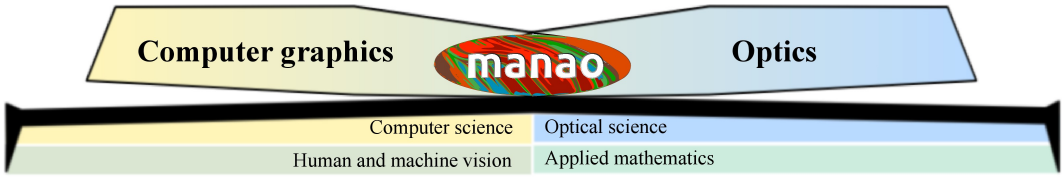


Figure 4. Related scientific domains of the MANAO project.

The *MANAO* project aims to study, acquire, model, and render 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 intersections 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 4) 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 [43] 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 [55], [56] and display [54] 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) and with the students issued from the “Institut d’Optique”, 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 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 (cf. Section 6.3). 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 [45] or differential analysis [74], 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, see Section 4.1) 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 Figure3 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 [5], 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 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 [51] in order to develop tailored strategies [88]. For instance, starting from simple direct lighting, more complex phenomena

have been progressively introduced: first diffuse indirect illumination [49], [81], then more generic inter-reflections [58], [43] and volumetric scattering [78], [40]. 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 [39]. 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 [80] but are difficult to extend to non-piecewise-constant data [83]. More recently, researches prefer the use of Spherical Radial Basis Functions [86] or Spherical Harmonics [73]. For more complex data, such as reflective properties (e.g., BRDF [68], [59] - 4D), ray-space (e.g., Light-Field [65] - 4D), spatially varying reflective properties (6D - [77]), new models, and representations are still investigated such as rational functions [19] or dedicated models [28] and parameterizations [79], [84]. 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 [19].

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 [74] and frequency analysis [45]). Such an approach has leaded us to exhibit the influence of surfaces flows (depth and normal gradients) into lighting pattern deformation (see Figure 8). For a **human observer**, this correspond to one recent trend in computer graphics that takes into account the human visual systems [47] 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.

For 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 [65], [54]. We consider projecting systems and surfaces [35], for personal use, virtual reality and augmented reality [31]. 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 [2], [48]. These resulting systems

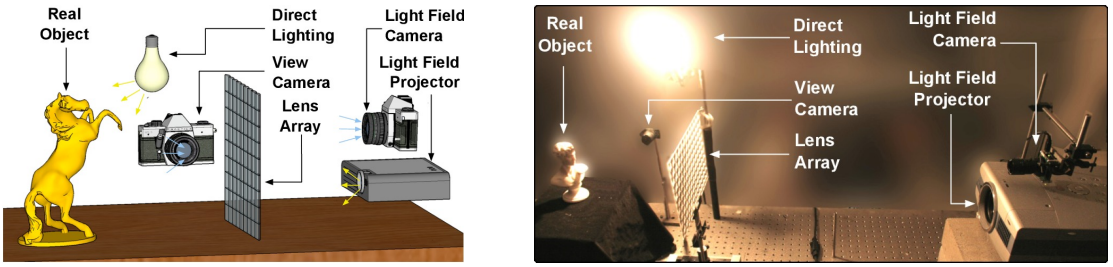


Figure 5. Light-Field transfer: global illumination between real and synthetic objects [38]

have to acquire the different phenomena described in Axis 1 and to display them, in an efficient manner [52], [29], [53], [56]. 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 [57].

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 [56], [72]). Similarly, increasing the number of viewpoints is a way toward the creation of full 3D displays [65]. 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. Furthermore, this leads to solutions that are not energy efficient and thus cannot be embedded into mobile devices. 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., [71], [89]).

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. [38]. 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 [63] are one of the key technologies to develop such acquisition (e.g., Light-Field camera [1] [57] and acquisition of light-sources [2]), projection systems (e.g., auto-stereoscopic displays). Such an approach is versatile and may be applied to improve classical optical instruments [62]. More generally, by designing unified optical and digital systems [69], 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

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

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 [19]. Second, we have to integrate signals from multiple sensors (such as GPS, accelerometer, ...) to prevent some computation (e.g., [17]). Finally, the experience of the group in surface modeling help the design of optical surfaces [60] 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 [70] (b) Shadows [30] (c) Shape enhancement [8] (d) Shape depiction [27]

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.

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 [9], [61] or stylized shading [33],[8] 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 [15] 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 [10], motion blur [46], depth of field [82], 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 [3], [44]. 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 [37]. Independent image processing for each viewpoint may disturb the feeling of depth by introducing inconsistent information in each images. Finally, more dedicated displays [54] 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.

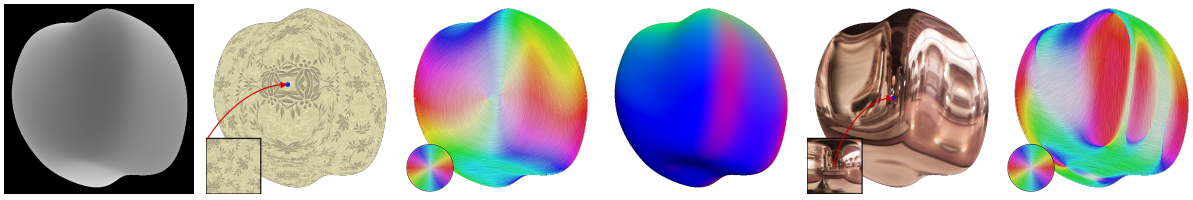
Results: High-level primitives and hybrid representations for appearance and shape.

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 [4], [1]. 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., [6]): 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 [6]), 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 [21] (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 Team

3. Scientific Foundations

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.

Transfering 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 experiment of collaboration with industrial partners. These collaborations bring us new problems to solve, with short-term or medium-term transfert 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.

MIMETIC Team

3. Scientific Foundations

3.1. Biomechanics and Motion Control

Human motion control is a very 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 exist 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 infinity of muscle activation patterns for a unique joint rotation. From the reactive point of view there exist 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.

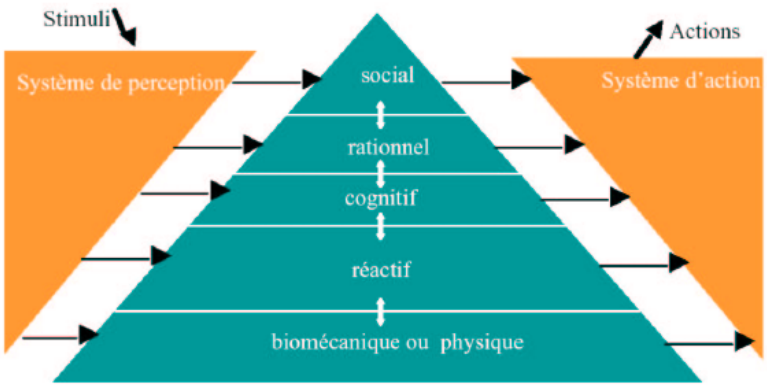


Figure 3. Layers of the motion control natural system in humans.

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 exist infinity of muscle activation patterns, especially when dealing with antagonist muscles. An unresolved problem is to define how using the above criteria to retrieve the actual activation patterns while optimization approaches still lead to unrealistic ones. It is still an open problem that will require multidisciplinary skills including computer simulation, constraint solving, biomechanics, optimal control, physiology and neurosciences.

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 combination 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 plan whose angle in the state space is associated with energy expenditure. Although there exists knowledge on specific motion, 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 objectivly 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.

Neurosciences also address 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 big 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 enhanche 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 knowleged in human motion control.

3.2. Experiments in Virtual Reality

Understanding interaction between humans is very 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, it is thus very complex to understand the influence of each of them on the interaction. 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 repeatible 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 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 observe its influence on the perception of the immersed subject. VR can then be used to understand what information are picked up to make a decision. Moreover, cut-offs can also be used to obtain temporal information about when these information are picked up. When the subject can moreover react as in 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 the 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. Computational geometry

Computational geometry is a branch of computer science devoted to the study of algorithms which can be stated in terms of geometry. It aims at studying algorithms for combinatorial, topological and metric problems concerning sets of points in Euclidian spaces. Combinatorial computational geometry focuses on three main problem classes: static problems, geometric query problems and dynamic problems.

In static problems, some input is given and the corresponding output needs to be constructed or found. Such problems include linear programming, Delaunay triangulations, and Euclidian shortest paths for instance. In geometric query problems, commonly known as geometric search problems, the input consists of two parts: the search space part and the query part, which varies over the problem instances. The search space typically needs to be preprocessed, in a way that multiple queries can be answered efficiently. Some typical problems are range searching, point location in a portioned space, nearest neighbor queries for instance. In dynamic problems, the goal is to find an efficient algorithm for finding a solution repeatedly after each incremental modification of the input data (addition, deletion or motion of input geometric elements). Algorithms for problems of this type typically involve dynamic data structures. Both of previous problem types can be converted into a dynamic problem, for instance, maintaining a Delaunay triangulation between moving points.

The Mimetic team works on problems such as crowd simulation, spatial analysis, path and motion planning in static and dynamic environments, camera planning with visibility constraints for instance. The core of those problems, by nature, relies on problems and techniques belonging to computational geometry. Proposed models pay attention to algorithms complexity to propose models compatible with performance constraints imposed by interactive applications.

MINT Project-Team

3. Scientific Foundations

3.1. Human-Computer Interaction

The scientific approach that we follow considers user interfaces as means, not an end: our focus is not on interfaces, but on interaction considered as a phenomenon between a person and a computing system [30]. We *observe* this phenomenon in order to understand it, i.e. *describe* it and possibly *explain* it, and we look for ways to significantly *improve* it. HCI borrows its methods from various disciplines, including Computer Science, Psychology, Ethnography and Design. Participatory design methods can help determine users' problems and needs and generate new ideas, for example [35]. Rapid and iterative prototyping techniques allow to decide between alternative solutions [31]. Controlled studies based on experimental or quasi-experimental designs can then be used to evaluate the chosen solutions [37]. One of the main difficulties of HCI research is the doubly changing nature of the studied phenomenon: people can both adapt to the system and at the same time adapt it for their own specific purposes [34]. As these purposes are usually difficult to anticipate, we regularly *create* new versions of the systems we develop to take into account new theoretical and empirical knowledge. We also seek to *integrate* this knowledge in theoretical frameworks and software tools to disseminate it.

3.2. Numerical and algorithmic real-time gesture analysis

Whatever is the interface, user provides some curves, defined over time, to the application. The curves constitute a gesture (positionnal information, yet may also include pressure). Depending on the hardware input, such a gesture may be either continuous (e.g. data-glove), or not (e.g. multi-touch screens). User gesture can be multi-variate (several fingers captured at the same time, combined into a single gesture, possibly involving two hands, maybe more in the context of co-located collaboration), that we would like, at higher-level, to be structured in time from simple elements in order to create specific command combinations.

One of the scientific fundations of the research project is an algorithmic and numerical study of gesture, which we classify into three points:

- *clustering*, that takes into account intrinsic structure of gesture (multi-finger/multi-hand/multi-user aspects), as a lower-level treatment for further use of gesture by application;
- *recognition*, that identifies some semantic from gesture, that can be further used for application control (as command input). We consider in this topic multi-finger gestures, two-handed gestures, gesture for collaboration, on which very few has been done so far to our knowledge. On the contrary, in the case of single gesture case (i.e. one single point moving over time in a continuous manner), numerous studies have been proposed in the current literature, and interestingly, are of interest in several communities: HMM [38], Dynamic Time Warping [40] are well-known methods for computer-vision community, and hand-writing recognition. In the computer graphics community, statistical classification using geometric descriptors has previously been used [36]; in the Human-Computer interaction community, some simple (and easy to implement) methods have been proposed, that provide a very good compromise between technical complexity and practical efficiency [39].
- *mapping to application*, that studies how to link gesture inputs to application. This ranges from transfer function that is classically involved in pointing tasks [32], to the question to know how to link gesture analysis and recognition to the algorithmic of application content, with specific reference examples.

We ground our activity on the topic of numerical algorithm, expertise that has been previously achieved by team members in the physical simulation community (within which we think that aspects such as elastic deformation energies evaluation, simulation of rigid bodies composed of unstructured particles, constraint-based animation... will bring up interesting and novel insights within HCI community).

3.3. Design and control of haptic devices

Our scientific approach in the design and control of haptic devices is focused on the interaction forces between the user and the device. We search of controlling them, as precisely as possible. This leads to different designs compared to other systems which control the deformation instead. The research is carried out in three steps:

- *identification:* we measure the forces which occur during the exploration of a real object, for example a surface for tactile purposes. We then analyze the record to deduce the key components – *on user's point of view* – of the interaction forces.
- *design:* we propose new designs of haptic devices, based on our knowledge of the key components of the interaction forces. For example, coupling tactile and kinesthetic feedback is a promising design to achieve a good simulation of actual surfaces. Our goal is to find designs which leads to compact systems, and which can stand close to a computer in a desktop environment.
- *control:* we have to supply the device with the good electrical conditions to accurately output the good forces.

POTIOC Team

3. Scientific Foundations

3.1. Introduction

The design of new user interfaces is a complex process that requires tackling research challenges at different levels. First, at a technological level, the input and output interaction space is becoming richer and richer. We will explore the new input/output modalities offered by such a technological evolution, and we will contribute to extend these modalities for the purpose of our main objective, which is to make 3D digital worlds available to all. Then, we will concentrate on the design of good interaction techniques that rely on such input/output modalities, and that are dedicated to the population targeted by this project, i.e. general public, specialists which are not 3D experts, and people with impairments. Finally, a large part of our work will be dedicated to the understanding and the assessment of user interaction. In particular, we will conduct user studies to guide the design of hardware and software UI, to evaluate them, and to better understand how a user interacts with 3D environments.

These three levels, input/output modalities, interaction techniques, and human factors will be the three main research directions of Potioc. Of course, they are extremely linked, and they cannot be studied independently, one after the other. In particular, user studies will follow the design process of hardware/software user interfaces from the beginning to the end, and both hardware and software exploration will be interdependent. The design of a new 3D user interface will thus require some work at different levels, as illustrated in Figure 2 . All members of Potioc will contribute in each of these research directions.

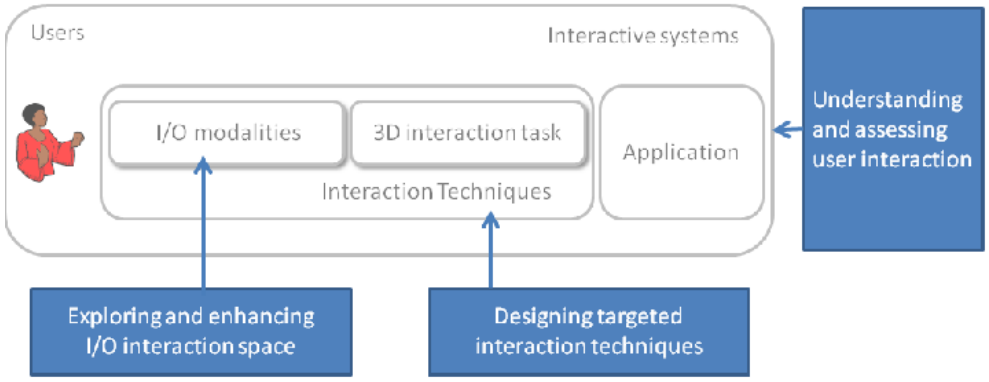


Figure 2. Diagram of an interactive system and the three main research axes of the Potioc project (blue boxes).

3.2. Exploring and enhancing input/output interaction space

The Potioc project-team will be widely oriented towards new innovative input and output modalities, even if standard approaches based on keyboard/mouse and standard screens will not be excluded. This includes motor-based interfaces, and physiological interfaces like BCI, as well as stereoscopic display and augmented reality setups. These technologies may have a great potential for opening 3D digital worlds to anyone, if they are correctly exploited.

We will explore various input/output modalities. Of course, we will not explore all of them at the same time, but we do not want to set an agenda either, for focusing on one of them. For a given need fed by end-users, we will choose among the various input/output modalities the ones that have the biggest potential. In the following paragraphs, we explain in more details the research challenges we will focus on to benefit from the existing and upcoming technologies.

3.2.1. Real-time acquisition and signal processing

There is a wide number of sensors that can detect users' activity. Beyond the mouse that detects x and y movements in a plane, various sensors are dedicated to the detection of 3D movements, pressure, brain and physiological activity, and so on. These sensors provide information that may be very rich, either to detect command intent from the user or to estimate and understand the user's state in real-time, but that are difficutly exploitable as it. Hence, a major challenge here is to extract the relevant information from the noisy raw data provided by the sensor.

An example, and important research topic in Potioc, is in the analysis of brain signals for the design of BCI. Indeed, brain signals are usually measured by EEG, such EEG signals being very noisy, complex and non-stationary. Moreover, for BCI-based applications, they need to be processed and analyzed in real-time. Finally, EEG signals exhibit large inter-user differences and there are usually few examples of EEG signals available to tune the BCI to a given user (we cannot ask the user to perform thousands of time the same mental task just to collect examples). As such, appropriate signal processing algorithms must be designed in order to robustly identify EEG patterns reflecting the user's intention. The research challenges are thus to design algorithms with high performance (in terms of rate of correctly identify user's state) anytime, anywhere, that are fully automatic and with minimal or no calibration time. In other words, we must design BCI that are convenient, comfortable and efficient enough so that they can be accepted and used by the end-user. Indeed, most users, in particular healthy users in the general public are used to highly convenient and efficient input devices (e.g., a simple mouse) and would not easily tolerate systems with a lower performance. Achieving this would make BCI good enough to be usable outside laboratories, e.g., for video gamers or patients. This will also make BCI valuable and reliable evaluation tools, e.g., to understand users' state during a given task. To address these challenges, pattern recognition and machine learning techniques are often used in order to find the optimal signal processing parameters. Similar approaches may contribute to the analysis of signals coming from other input devices than BCI. An example is the exploitation of depth cameras, where we need to find relevant information from noisy signals. Other emerging technologies will require similar attention, where the goal will be to transform an unstructured raw signal into a set of higher level descriptors that can be used as input parameters for controlling interaction techniques.

3.2.2. Restitution and perceptive feedback

Similarly to the input side, the feedback provided to the user through various output modalities will be explored in Potioc. Beyond the standard screens that are commonly used, we will explore various displays. In particular, in the scope of visual restitution, we will notably focus on large screens and tables, mobile setups and projection on real objects, and stereoscopic visualization. The challenge here will be to conceive good visual metaphors dedicated to these unconventional output devices in order to maximize the attractiveness and the pleasure linked to the use of these technologies.

For example, we will investigate the use of stereoscopic displays for extending the current visualization approaches. Indeed, stereoscopic visualization has been little explored outside the complex VR setups dedicated to professional users. We believe that this modality may be very interesting for non-expert users, in wider contexts. To reach this goal, we will thus concentrate on new visual metaphors that benefit from stereoscopic visualization, and we will explore how, when, and where stereoscopy may be used.

Depending on the targeted interaction tasks, we may also investigate various additional output modalities such as tangible interaction, audio displays, and so on. In any case, our approach will be the same, which is understanding how new perceptive modalities may push the frontier of our current interactive systems.

3.2.3. Creation of new systems

In addition to the exploration and the exploitation of existing input and output modalities for enhancing interaction with 3D content, we may also contribute to extend the current input/output interaction space by building new interactive systems. This will be done by combining hardware components, or by collaborating with mechanics/electronics specialists.

3.3. Designing targeted interaction techniques

In the previous section, we focused on the input/output interaction space, which is closely related to hardware components. In this part, we focus on the design of interaction techniques, which we define here as the mean through which a user will complete an interaction task from a given interaction space. Even if this is naturally also linked to the underlying hardware components, the research conducted in this axis of the project will mainly concern software developments.

Similar to the input/output interaction space, the design of interaction techniques requires focusing on both the motor and the sensory components. In our 3D spatial context, thus the challenges will be to find good mappings between the available input and the DOF that need to be controlled in the 3D environment, and to provide relevant feedback to users so that they can understand well what they are doing.

The design of interaction techniques should be strongly guided by the targeted end-users. For example, a 3D UI dedicated to an expert user will not suit a novice user, and the converse is also true. In Potioc, where the final goal is to open 3D digital worlds to anyone, we will concentrate on the general public, specialists that are not 3D experts, and people with impairments.

3.3.1. General public

3D UIs have mainly been designed for professional use. For example, modeling tools require expertize to be used correctly and, consequently, they exclude the general public from the process of creating 3D content. Similarly, immersive technologies have been dedicated to professional users for a long time. Therefore, immersive 3D interaction techniques have generally been thought for trained users, and they may not fit well with a general public context. In Potioc, an important motivation will be to re-invent 3D UIs to adapt them to the general public. This motivation will guide us towards new approaches that have been little explored until now. In particular, to reach our objective, we will give a strong importance to the following criteria:

- Intuitiveness: a very short learning curve will be required.
- Enjoyability: this is needed to motivate novice users in the complex process of interaction with 3D content.
- Robustness: the UIs should support untrained users that may potentially interact with unpredictable actions.

In addition, we will keep connected with societal and technological factors surrounding the general public. For example, [multi]touch-screens have become very popular these past few years, and everyone tend to be familiar with a standard gesture vocabulary (e.g. pinch gestures and flicking gestures). We will rely on these commonly acquired *way-of-interact* to optimize the acceptability of the 3D UIs we will design. In this part of the project the challenge will be to conceive 3D UIs that offer a high degree of interactivity, while ensuring an easy access to technology, as well as a wide adherence.

3.3.2. Specialists

General public will be one of the main targets of Potioc for the design of 3D UIs. However, we do not exclude specialists, who have little experience with 3D interaction. These specialists can be for example artists, archaeologists, or architects. In any case, we are convinced that 3D digital worlds could benefit to such categories of users if we propose dedicated 3D UIs that allows them to better understand, communicate, or create, with their respective skills. Because such specialists will gain expertize while interacting with 3D content, it will be necessary to design 3D UIs that can adapt to their evolving level of expertize. In particular, the UIs should be easy to use and attractive enough to encourage new users. At the same time, they should provide advanced features that the specialist can discover while gaining expertize.

3.3.3. People with impairments

While the general public has been only scarcely considered as a potential target audience for 3D digital worlds, another category of users is even more neglected: people with impairments. Indeed, such people, in particular those with motor impairments, are unable to use classical input devices, since they have been designed for healthy users. People with motor impairment have to use dedicated input devices, adapted to their disabilities, such as a single switch. Since such input devices usually have much fewer degrees of freedom than classical devices, it is necessary to come up with appropriate interaction techniques in order to efficiently use this limited number of DOF to still enable the user to perform complex tasks in the 3D environment. In Potioc, our focus will be on the use of BCI to enable motor impaired users to interact with 3D environment for learning, creation and entertainment. Indeed, BCI enable a user to interact without any motor movement.

3.4. Understanding and assessing user interaction

The exploration of the input/output interaction space, and the design of new interaction techniques, are strongly linked with human factors, which will be the third research axis of the Potioc project. Indeed, to guide the developments described in the previous sections, we first need to well understand users' motor and cognitive skills for the completion of 3D interaction tasks. This will be explored thanks to *a-priori* experiments. In order to evaluate our hardware and software interfaces, we will conduct *a-posteriori* user studies. Finally, we will explore new approaches for a real-time cognitive analysis of the performance and the experience of a user interacting with a 3D environment.

The main challenge in this part of the project will be to design good experimental protocols that will allow us to finely analyze various parameters for improving our interfaces. In 2D, there exist many standard protocols and prediction laws for evaluating UIs (e.g. Fitts law and ISO 9241). This is not the case in 3D. Consequently, a special care must be taken when evaluating interaction in 3D spatial contexts.

In addition to the standard experiments we will conduct in our lab, we will conduct large scale experiments thanks to the strong collaboration we have with the center for the widespread of scientific culture, Cap Sciences (see Collaboration section). With such kind of experiments, we will be able to test hundreds of participants, with various ages, gender, or level of expertize that we will be able to track thanks to the Navinum system [3], and this during long period of time. A challenge for us will be to gain benefit from this wealth of information for the development of our 3D UIs.

3.4.1. A-priori user studies

Before designing 3D UIs, it is important to understand what a user is good at, and what may cause difficulties. This is true at a motor level, as well as a cognitive level. For example, are users able to coordinate the movements of several fingers on a touchscreen at the same time, or are they able to finely control the quantity of force applied on it while moving their hand? Similarly, are the users able to mentally predict a 3D rotation, and how many levels of depth are they able to distinguish when visualizing stereoscopic images? To answer these questions, we will conduct preliminary studies.

Our research in that direction will guide our developments for the other research axes described above. For example, it will be interesting to explore touch-based 3D UIs that take into account several level of force if we see that this parameter can be easily handled by users. On the other hand, if the results of a-priori tests show that this input cannot be easily controlled, then we will not push forward that direction.

The members of Potioc have already conducted such kinds of experiments, and we will continue our work in that direction. For some investigations, we will collaborate with psychologists and experts in cognitive sciences (see Collaborations section) to explore in more depth motor and cognitive human skills.

[3] Navinum is a system based on a RFID technology that is used to collect informations about the activity of the visitors in Cap Sciences. http://www.scribd.com/doc/55178878/Dossier-de-Presse-Numerique-100511

A-priori studies will allow us to understand how users tend to "*naturally*" interact to complete 3D interaction tasks, and to understand which feedback are the best suited. This will be a first answer to our global quest of providing pleasant interfaces. Indeed, this will allow us to adapt the UIs to the users, and not the opposite. This should enhance the global acceptability and motivation of users facing a new interactive system.

3.4.2. A-posteriori user studies

In Potioc, we will conceive new hardware and software interfaces. To validate these UIs, and to improve them, we will conduct user experiments, as classically done in the field of HCI. This is a standard methodology that we currently follow (see Bibliography). We will do this in our lab, and in Cap Sciences.

Beyond the standard evaluation criteria that are based on performance for speed, accuracy, coordination, and so on, we will also consider other criteria that are more relevant for the Potioc project. Indeed, we will give a great importance to enjoyability, pleasure of use, accessibility, and so on. Consequently, we will need to redefine the standard way to evaluate UIs. Once again, our relationship with Cap Sciences will help us in such investigations. The use of questionnaires will be a way to better understand how an interface should be designed to reach a successful use. In addition, we will observe and analyze how visitors tend to interact with various interfaces we will propose. For example, we will collect information like the time spent on a given interactive system or the number of smiles recorded during an interaction process. The identification of good criteria to use for the evaluation of a popular 3D UI will be one of the research directions of our team.

Conducting such *a-posteriori* studies, in particular with experts of mediation, with new criteria of success, will be a second answer to our goal of evaluating the pleasure linked to the use of 3D UIs.

3.4.3. Real-time cognitive analysis

Classically, the user's subjective preferences for a given 3DUI are assessed using questionnaires. While these questionnaires provide important information, this is only a partial, biased, a-posteriori/a-priori measure, since they are collected before or after the 3D interaction process. When questionnaires are administered during 3D interaction, this interrupts and disturbs the user, hence biasing the evaluation. Moreover, while evaluating performance and usefulness is now well described and understood, evaluating the user's experiance and thus the system usability appears as much more difficult, with a lack of systematic and standard approaches. Ideally, we would like to measure the user response and subjective experience while he/she is using the 3DUI, i.e., in real-time and without interrupting him/her, in order to precisely identify the UI pros and cons. Questionnaires cannot provide such a measure.

Fortunately, it has been recently shown that BCI could be used in a passive way, to monitor the user's mental state. More precisely, recent results suggested that appropriately processed EEG signals could provide information about mental states such as error perception, attention or mental workload. As such, BCI are emerging as a new tool to monitor a user's mental state and brain responses to various stimuli, in real-time. In the Potioc project, we propose a completely new way to evaluate 3DUI: rather than relying only on questionnaires to estimate the user's subjective experience, we propose to exploit passive BCI to estimate the user's mental state in real-time, without interrupting nor disturbing him or her, while he/she is using the 3DUI. In particular, we aim at measuring and processing EEG and other biosignals (e.g., pulse, galvanic skin response, electromyogram) in real-time in order to estimate mental states such as interaction error potentials or workload/attention levels, among others. This will be used to finely identify how intuitive, easy-to-use and (ideally) enjoyable any given 3D UI is. More specifically, it will allow us to identify how, when and where the UI has flaws. Because the analysis will occur in real-time, we will potentially be able to modify the interface while the user is interacting. This should lead to a better understanding of 3D interaction. The work that will be achieved in this area could potentially also be useful for 2D interface design. However, since Potioc's main target is 3DUI, we will naturally focus the real-time cognitive evaluations on 3D contexts, with specific targets such as depth perception, or perception of 3D rotations.

This real-time cognitive analysis will be a third answer to reach the objectives of Potioc, which are to open 3D digital worlds to everyone by increasing the pleasure of use.

REVES Project-Team

3. Scientific Foundations

3.1. Rendering

We consider plausible rendering to be a first promising research direction, both for images and for sound. Recent developments, such as point rendering, image-based modeling and rendering, and work on the simulation of aging indicate high potential for the development of techniques which render *plausible* rather than extremely accurate images. In particular, such approaches can result in more efficient renderings of very complex scenes (such as outdoors environments). This is true both for visual (image) and sound rendering. In the case of images, such techniques are naturally related to image- or point-based methods. It is important to note that these models are becoming more and more important in the context of network or heterogeneous rendering, where the traditional polygon-based approach is rapidly reaching its limits. Another research direction of interest is realistic rendering using simulation methods, both for images and sound. In some cases, research in these domains has reached a certain level of maturity, for example in the case of lighting and global illumination. For some of these domains, we investigate the possibility of technology transfer with appropriate partners. Nonetheless, certain aspects of these research domains, such as visibility or high-quality sound still have numerous and interesting remaining research challenges.

3.1.1. Plausible Rendering

3.1.1.1. Alternative representations for complex geometry

The key elements required to obtain visually rich simulations, are sufficient geometric detail, textures and lighting effects. A variety of algorithms exist to achieve these goals, for example displacement mapping, that is the displacement of a surface by a function or a series of functions, which are often generated stochastically. With such methods, it is possible to generate convincing representations of terrains or mountains, or of non-smooth objects such as rocks. Traditional approaches used to represent such objects require a very large number of polygons, resulting in slow rendering rates. Much more efficient rendering can be achieved by using point or image based rendering, where the number of elements used for display is view- or image resolution-dependent, resulting in a significant decrease in geometric complexity. Such approaches have very high potential. For example, if all object can be rendered by points, it could be possible to achieve much higher quality local illumination or shading, using more sophisticated and expensive algorithms, since geometric complexity will be reduced. Such novel techniques could lead to a complete replacement of polygon-based rendering for complex scenes. A number of significant technical challenges remain to achieve such a goal, including sampling techniques which adapt well to shading and shadowing algorithms, the development of algorithms and data structures which are both fast and compact, and which can allow interactive or real-time rendering. The type of rendering platforms used, varying from the high-performance graphics workstation all the way to the PDA or mobile phone, is an additional consideration in the development of these structures and algorithms. Such approaches are clearly a suitable choice for network rendering, for games or the modelling of certain natural object or phenomena (such as vegetation, e.g. Figure 1 , or clouds). Other representations merit further research, such as image or video based rendering algorithms, or structures/algorithms such as the "render cache" [37], which we have developed in the past, or even volumetric methods. We will take into account considerations related to heterogeneous rendering platforms, network rendering, and the appropriate choices depending on bandwith or application. Point- or image-based representations can also lead to novel solutions for capturing and representing real objects. By combining real images, sampling techniques and borrowing techniques from other domains (e.g., computer vision, volumetric imaging, tomography etc.) we hope to develop representations of complex natural objects which will allow rapid rendering. Such approaches are closely related to texture synthesis and image-based modeling. We believe that such methods will not replace 3D (laser or range-finger) scans, but could be complementary, and represent a simpler and lower cost alternative for certain applications (architecture, archeology etc.). We are also investigating methods for

adding "natural appearance" to synthetic objects. Such approaches include *weathering* or *aging* techniques, based on physical simulations [27], but also simpler methods such as accessibility maps [34]. The approaches we intend to investigate will attempt to both combine and simplify existing techniques, or develop novel approaches founded on generative models based on observation of the real world.

3.1.1.2. Plausible audio rendering

Similar to image rendering, plausible approaches can be designed for audio rendering. For instance, the complexity of rendering high order reflections of sound waves makes current geometrical approaches inappropriate. However, such high order reflections drive our auditory perception of "reverberation" in a virtual environment and are thus a key aspect of a plausible audio rendering approach. In complex environments, such as cities, with a high geometrical complexity, hundreds or thousands of pedestrians and vehicles, the acoustic field is extremely rich. Here again, current geometrical approaches cannot be used due to the overwhelming number of sound sources to process. We study approaches for statistical modeling of sound scenes to efficiently deal with such complex environments. We also study perceptual approaches to audio rendering which can result in high efficiency rendering algorithms while preserving visual-auditory consistency if required.

Figure 1. Plausible rendering of an outdoors scene containing points, lines and polygons [26], representing a scene with trees, grass and flowers. We can achieve 7-8 frames per second compared to tens of seconds per image using standard polygonal rendering.

3.1.2. High Quality Rendering Using Simulation

3.1.2.1. Non-diffuse lighting

A large body of global illumination research has concentrated on finite element methods for the simulation of the diffuse component and stochastic methods for the non-diffuse component. Mesh-based finite element approaches have a number of limitations, in terms of finding appropriate meshing strategies and form-factor calculations. Error analysis methodologies for finite element and stochastic methods have been very different in the past, and a unified approach would clearly be interesting. Efficient rendering, which is a major advantage of finite element approaches, remains an overall goal for all general global illumination research. For certain cases, stochastic methods can be efficient for all types of light transfers, in particular if we require a view-dependent solution. We are also interested both in *pure* stochastic methods, which do not use finite element techniques. Interesting future directions include filtering for improvement of final image quality as well as beam tracing type approaches [35] which have been recently developed for sound research.

3.1.2.2. Visibility and Shadows

Visibility calculations are central to all global illumination simulations, as well as for all rendering algorithms of images and sound. We have investigated various global visibility structures, and developed robust solutions for scenes typically used in computer graphics. Such analytical data structures [31], [30], [29] typically have robustness or memory consumption problems which make them difficult to apply to scenes of realistic size. Our solutions to date are based on general and flexible formalisms which describe all visibility event in terms of generators (vertices and edges); this approach has been published in the past [28]. Lazy evaluation, as well as hierarchical solutions, are clearly interesting avenues of research, although are probably quite application dependent.

3.1.2.3. Radiosity

For purely diffuse scenes, the radiosity algorithm remains one of the most well-adapted solutions. This area has reached a certain level of maturity, and many of the remaining problems are more technology-transfer oriented We are interested in interactive or real-time renderings of global illumination simulations for very complex scenes, the "cleanup" of input data, the use of application-dependent semantic information and mixed representations and their management. Hierarchical radiosity can also be applied to sound, and the ideas used in clustering methods for lighting can be applied to sound.

3.1.2.4. High-quality audio rendering

Our research on high quality audio rendering is focused on developing efficient algorithms for simulations of geometrical acoustics. It is necessary to develop techniques that can deal with complex scenes, introducing efficient algorithms and data structures (for instance, beam-trees [32] [35]), especially to model early reflections or diffractions from the objects in the environment. Validation of the algorithms is also a key aspect that is necessary in order to determine important acoustical phenomena, mandatory in order to obtain a high-quality result. Recent work by Nicolas Tsingos at Bell Labs [33] has shown that geometrical approaches can lead to high quality modeling of sound reflection and diffraction in a virtual environment (Figure 2). We will pursue this research further, for instance by dealing with more complex geometry (e.g., concert hall, entire building floors).

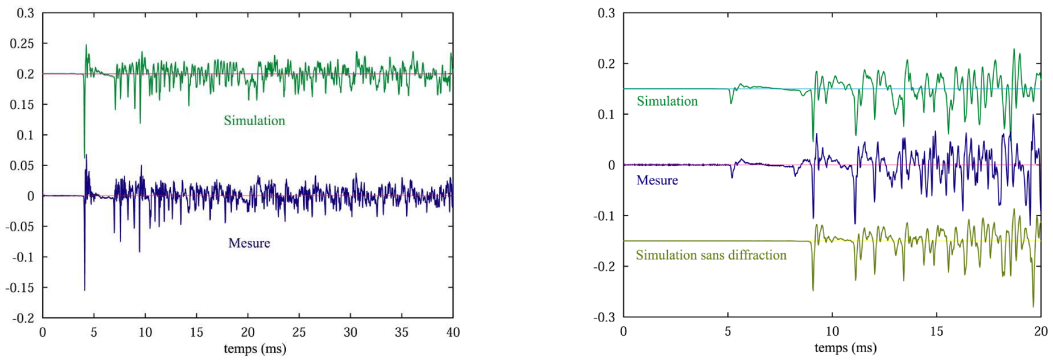


Figure 2. A comparison between a measurement (left) of the sound pressure in a given location of the "Bell Labs Box", a simple test environment built at Bell Laboratories, and a high-quality simulation based on a beam-tracing engine (right). Simulations include effects of reflections off the walls and diffraction off a panel introduced in the room.

Finally, several signal processing issues remain in order to properly and efficiently restitute a 3D soundfield to the ears of the listener over a variety of systems (headphones, speakers). We would like to develop an open and general-purpose API for audio rendering applications. We already completed a preliminary version of a software library: AURELI [36].

VR4I Team

3. Scientific Foundations

3.1. Panorama

Our main concern is to allow real users to interact naturally within shared virtual environments as interaction can be the result of an individual interaction of one user with one object or a common interaction of several users on the same object. The long-term purpose of the project is to propose interaction modalities within virtual environments that bring **acting in Virtual Reality as natural as acting in reality**.

Complex physically based models have to be proposed to represent the virtual environment, complex multi-modal interaction models have to be proposed to represent natural activity and complex collaborative environments have to be proposed to ensure effective collaborative interactions.

The long term objectives of VR4i are:

- Improving the accuracy of the virtual environment representation for more interactivity and better perception of the environment;
- Improving the multi-modal interaction for more natural interactions and better perception of the activity;
- Improving the use of virtual environments for real activity and open to human science for evaluation and to engineering science for applications.

Thus, we propose three complementary research axes:

- Physical modeling and simulation of the environment
- Multimodal immersive interaction
- Collaborative work in Collaborative Virtual Environments (CVE)

3.2. Physical modeling and simulation

The first aspect is the modeling and the simulation of the virtual world that represents properly the physical behavior of the virtual world that sustains a natural interaction through the different devices. The main challenge is the search of the trade-off between accuracy and performance to allow effective manipulation, in interactive time, by the user. This trade-off is a key point while the user closes the interaction loop. Namely, the accuracy of the simulation drives the quality of the phenomenon to perceive and the performance drives the sensori-motor feelings of the user. Proposing new controlled algorithms for physical based simulation of the virtual world is certainly a key point for meeting this trade-off. We believe that the mechanical behavior of objects as to be more studied and to be as close as possible to their real behavior. The devices may act as a both way filter on the action and on the perception of the simulated world, but improving the representation of rigid objects submitted to contact, of deformable objects, of changing state object and of environments that include mixed rigid and deformable objects is needed in order to compute forces and positions that have a physical meaning. The interaction between tools and deformable objects is still a challenge in assembly applications and in medical applications. The activity of the user in interaction with the immersive environment will allow to provide method to qualify the quality of the environment and of the interaction by proposing a bio-mechanical user's Alter Ego. We believe that the analysis of the forces involved during an immersive activity will give us keys to design more acceptable environments. As the goal is to achieve more and more accurate simulation that will require more and more computation time, the coupling between physical modeling and related simulation algorithms is of first importance. Looking for genericity will ensure correct deployment on new advanced hardware platforms that we will use to ensure adapted performance. The main aim of this topic is to improve the simulation accuracy satisfying the simulation time constraints for improving the naturalness of interactions.

3.3. Multimodal immersive interaction

The second aspect concerns the design and evaluation of novel approaches for multimodal immersive interaction with virtual environments.

We aim at improving capabilities of selection and manipulation of virtual objects, as well as navigation in the virtual scene and control of the virtual application. We target a wide spectrum of sensory modalities and interfaces such as tangible devices, haptic interfaces (force-feedback, tactile feedback), visual interfaces (e.g., gaze tracking), locomotion and walking interfaces, and brain-computer interfaces. We consider this field as a strong scientific and technological challenge involving advanced user interfaces, but also as strongly related to user's perceptual experience. We promote a perception-based approach for multimodal interaction, based on collaborations with laboratories of the Perception and Neuroscience research community.

The introduction of a third dimension when interacting with a virtual environment makes inappropriate most of the classical techniques used successfully in the field of 2D interaction with desktop computers up to now. Thus, it becomes successfully used to design and evaluate new paradigms specifically oriented towards interaction within 3D virtual environments.

We aim at improving the immersion of VR users by offering them natural ways for navigation, interaction and application control, as these are the three main tasks within 3D virtual environments. Here we consider interactions as multimodal interactions, as described in the previous section. We also want to make the users forget their physical environment in benefit of the virtual environment that surrounds them and contribute to improve the feeling of immersion and of presence. To achieve this goal, we must ensure that users can avoid collisions with their surrounding real environment (the screens of the rendering system, the walls of the room) and can avoid lost of interaction tracking (keeping the user within the range of the physical interaction devices). To do that, we propose to take into account the surrounding real physical environment of the user and to include it in the virtual environment through a virtual representation. This explicit model of the real environment of the users will help users to forget it: throughout this model, the user will be aware (with visual, auditive or haptic feedback) of these virtual objects when he comes near their boundaries. We also have to investigate which physical limitations are the most important ones to perceive, and what are the best ways to make the users aware of their physical limitations.

3.4. Collaborative work in CVE's

The third aspect is to propose Collaborative Virtual Environments for several local or distant users. In these environments, distant experts could share their expertise for project review, for collaborative design or for analysis of data resulting from scientific computations in HPC context. Sharing the virtual environment is certainly a key point that leads to propose new software architectures ensuring the data distribution and the synchronization of the users.

In terms of interaction, new multi-modal interaction metaphors have to be proposed to tackle with the awareness of other users? activity. Here it is important to see a virtual representation of the other users, of their activity, and of the range of their action field, in order to better understand both their potential and their limitation for collaboration: what they can see, what they can reach, what their interaction tools are and which possibilities they offer.

Simultaneous collaborative interactions upon the same data through local representations of these data should be tackled by new generic algorithms dedicated to consistency management. Some solutions have to be proposed for distant collaboration, where it is not possible any more to share tangible devices to synchronize co-manipulation: we should offer some new haptic rendering to enforce users' coordination. Using physics engines for realistic interaction with virtual objects is also a challenge if we want to offer low latency feedback to the users. Indeed, the classical centralized approach for physics engines is not able to offer fast feedback to distant users, so this approach must be improved.

AXIS Project-Team (section vide)

DAHU Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

Dahu has strong connections with the Leo project-team in Saclay.

Dahu aims at developing mechanisms for high-level specifications of systems built around DBMS, that are easy to understand while also facilitating verification of critical properties. This requires developing tools that are suitable for reasoning about systems that manipulate data. Some tools for specifying and reasoning about data have already been studied independently by the database community and by the verification community, with various motivations. However, this work is still in its infancy and needs to be further developed and unified.

Most current proposals for reasoning about DBMS over XML documents are based on tree automata, taking advantage of the tree structure of XML documents. For this reason, the Dahu team is studying a variety of tree automata. This ranges from restrictions of "classical" tree automata in order to understand their expressive power, to extensions of tree automata in order to understand how to incorporate the manipulation of data.

Moreover, Dahu is also interested in logical frameworks that explicitly refer to data. Such logical frameworks can be used as high level declarative languages for specifying integrity constraints, format change during data exchange, web service functionalities and so on. Moreover, the same logical frameworks can be used to express the critical properties we wish to verify.

In order to achieve its goals, Dahu brings together world-class expertise in both databases and verification.

DREAM Project-Team

3. Scientific Foundations

3.1. Computer assisted monitoring and diagnosis of physical systems

keywords: monitoring,diagnosis, deep model, fault model, simulation, chronicle acquisition

Our work on monitoring and diagnosis relies on model-based approaches developed by the Artificial Intelligence community since the seminal studies by R. Reiter and J. de Kleer [63], [74]. Two main approaches have been proposed then: (i) the consistency-based approach, relying on a model of the expected correct behavior ; (ii) the abductive approach which relies on a model of the failures that might affect the system, and which identifies the failures or the faulty behavior explaining the anomalous observations. See the references [21], [23] for a detailed exposition of these investigations.

Since 1990, the researchers in the field have studied dynamic system monitoring and diagnosis, in a similar way as researchers in control theory do. What characterizes the AI approach is the use of qualitative models instead of quantitative ones and the importance given to the search for the actual source/causes of the faulty behavior. Model-based diagnosis approaches rely on qualitative simulation or on causal graphs in order to look for the causes of the observed deviations. The links between the two communities have been enforced, in particular for what concerns the work about discrete events systems and hybrid systems. Used formalisms are often similar (automata, Petri nets ,...) [28], [27].

Our team focuses on monitoring and on-line diagnosis of discrete events systems and in particular on monitoring by alarm management.

Two different methods have been studied by our team in the last years:

- In the first method, the automaton used as a model is transformed off-line into an automaton adapted to diagnosis. This automaton is called a *diagnoser*. This method has first been proposed by M. Sampath and colleagues [65]. The main drawback of this approach is its centralized nature that requires to explicitly build the global model of the system, which is most of the time unrealistic. It is why we proposed a decentralized approach in [60].
- In the second method, the idea is to associate each failure that we want to detect with a *chronicle* (or a scenario), i.e. a set of observable events interlinked by time constraints. The chronicle recognition approach consists in monitoring and diagnosing dynamic systems by recognizing those chronicles on-line [43], [62], [41].

  One of our research focus is to extend the chronicle recognition methods to a distributed context. Local chronicle bases and local recognizers are used to detect and diagnose each component. However, it is important to take into account the interaction model (messages exchanged by the components). Computing a global diagnosis requires then to check the synchronisation constraints between local diagnoses.

  Another issue is the chronicle base acquisition. An expert is often needed to create the chronicle base, and that makes the creation and the maintenance of the base very expensive. That is why we are working on an automatic method to acquire the base.

Developing diagnosis methodologies is not enough, especially when on-line monitoring is required. Two related concerns must be tackled, and are the topics of current research in the team:

- The ultimate goal is usually not merely to diagnose, but to put the system back in some acceptable state after the occurrence of a fault. One of our aim is to develop self-healable systems able to self-diagnose and -repair.

- When designing a system and equipping it with diagnosis capabilities, it may be crucial to be able to check off-line that the system will behave correctly, i.e., that the system is actually 'diagnosable'. A lot of techniques have been developed in the past (see Lafortune and colleagues [64]), essentially in automata models. We extended them to cope with temporal patterns. A recent focus has been to study the self-healability of systems (ability to self-diagnose and -repair).

3.2. Machine learning and data mining

keywords: machine learning, Inductive Logic Programming (ILP), temporal data mining, temporal abstraction, data-streams

The machine learning and data mining techniques investigated in the group aim at acquiring and improving models automatically. They belong to the field of machine or artificial learning [38]. In this domain, the goal is the induction or the discovery of hidden objects characterizations from their descriptions by a set of features or attributes. For several years we investigated Inductive Logic Programming (ILP) but now we are also working on data-mining techniques.

We are especially interested in structural learning which aims at making explicit dependencies among data where such links are not known. The relational (temporal or spatial) dimension is of particular importance in applications we are dealing with, such as process monitoring in health-care, environment or telecommunications. Being strongly related to the dynamics of the observed processes, attributes related to temporal or spatial information must be treated in a special way. Additionally, we consider that the legibility of the learned results is of crucial importance as domain experts must be able to evaluate and assess these results.

The discovery of spatial patterns or temporal relations in sequences of events involve two main steps: the choice of a data representation and the choice of a learning technique.

We are mainly interested in symbolic supervised and unsupervised learning methods. Furthermore, we are investigating methods that can cope with temporal or spatial relationships in data. In the sequel, we will give some details about relational learning, relational data-mining and data streams mining.

3.2.1. Relational learning

) Relational learning, also called inductive logic programming (ILP), lies at the intersection of machine learning, logic programming and automated deduction. Relational learning aims at inducing classification or prediction rules from examples and from domain knowledge. As relational learning relies on first order logic, it provides a very expressive and powerful language for representing learning hypotheses especially those learnt from temporal data. Furthermore, domain knowledge represented in the same language can also be used. This is a very interesting feature which enables taking into account already available knowledge and avoids starting learning from scratch.

Concerning temporal data, our work is more concerned with applying relational learning rather than developing or improving the techniques. Nevertheless, as noticed by Page and Srinivasan [59], the target application domains (such as signal processing in health-care) can benefit from adapting relational learning scheme to the particular features of the application data. Therefore, relational learning makes use of constraint programming to infer numerical values efficiently [66]. Extensions, such as QSIM [49], have also been used for learning a model of the behavior of a dynamic system [44]. Precisely, we investigate how to associate temporal abstraction methods to learning and to chronicle recognition. We are also interested in constraint clause induction, particularly for managing temporal aspects. In this setting, the representation of temporal phenomena uses specific variables managed by a constraint system [61] in order to deal efficiently with the associated computations (such as the covering tests).

For environmental data, we have investigated tree structures where a set of attributes describe nodes. Our goal is to find patterns expressed as sub-trees [37] with attribute selectors associated to nodes.

3.2.2. Data mining

Data mining is an unsupervised learning method which aims at discovering interesting knowledge from data. Association rule extraction is one of the most popular approach and has deserved a lot of interest in the last 10 years. For instance, many enhancements have been proposed to the well-known Apriori algorithm [25]. It is based on a level-wise generation of candidate patterns and on efficient candidate pruning having a sufficient relevance, usually related to the frequency of the candidate pattern in the data-set (i.e., the support): the most frequent patterns should be the most interesting. Later, Agrawal and Srikant proposed a framework for "mining sequential patterns" [26], which extends Apriori by coping with the order of elements in patterns.

In [54], Mannila and Toivonen extended the work of Aggrawal et al. by introducing an algorithm for mining patterns involving temporal episodes with a distinction between parallel and sequential event patterns. Later, in [42], Dousson and Vu Duong introduced an algorithm for mining chronicles. Chronicles are sets of events associated with temporal constraints on their occurrences. They generalize the temporal patterns of Mannila and Toivonen. The candidate generation is an Apriori-like algorithm. The chronicle recognizer CRS [40] is used to compute the support of patterns. Then, the temporal constraints are computed as an interval whose bounds are the minimal and the maximal temporal extent of the delay separating the occurrences of two given events in the data-set. Chronicles are very interesting because they can model a system behavior with sufficient precision to compute fine diagnoses. Their extraction from a data-set is reasonably efficient. They can be efficiently recognized on an input data stream.

Relational data-mining [22] can be seen as generalizing these works to first order patterns. In this field, the work of Dehaspe for extracting first-order association rules have strong links with chronicles. Another interesting research concerns inductive databases which aim at giving a theoretical and logical framework to data-mining [50], [39]. In this view, the mining process means to query a database containing raw data as well as patterns that are implicitly coded in the data. The answer to a query is, either the solution patterns that are already present in the database, or computed by a mining algorithm, e.g., Apriori. The original work concerns sequential patterns only [53]. We have investigated an extension of inductive database where patterns are very close to chronicles [69].

3.2.3. Mining data streams

) During the last years, a new challenge has appeared in the data mining community: mining from data streams [24]. Data coming for example from monitoring systems observing patients or from telecommunication systems arrive in such huge volumes that they cannot be stored in totality for further processing: the key feature is that “you get only one look at the data” [46]. Many investigations have been made to adapt existing mining algorithms to this particular context or to propose new solutions: for example, methods for building synopses of past data in the form of or summaries have been proposed, as well as representation models taking advantage of the most recent data. Sequential pattern stream mining is still an issue [55]. At present, research topics such as, sampling, summarizing, clustering and mining data streams are actively investigated.

A major issue in data streams is to take into account the dynamics of process generating data, i.e., the underlying model is evolving and, so, the extracted patterns have to be adapted constantly. This feature, known as *concept drift* [71], [51], occurs within an evolving system when the state of some hidden system variables changes. This is the source of important challenges for data stream mining [45] because it is impossible to store all the data for off-line processing or learning. Thus, changes must be detected on-line and the current mined models must be updated on line as well.

EXMO Project-Team

3. Scientific Foundations

3.1. Knowledge representation semantics

We usually work with semantically defined knowledge representation languages (like description logics [28], conceptual graphs and object-based languages). Their semantics is usually defined within model theory initially developed for logics. The languages dedicated to the semantic web (RDF and OWL) follow that approach. RDF is a knowledge representation language dedicated to the annotation of resources within the framework of the semantic web. OWL is designed for expressing ontologies: it describes concepts and relations that can be used within RDF.

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 is also called an ontology. An interpretation function (I) is inductively defined over the structure of the language to a structure called interpretation domain (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 these expressions. An expression (δ) is then a consequence of a set of expressions (o) if it is satisfied by all of their models (noted $o \models \delta$).

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. However, depending on the language and its semantics, the decidability, i.e., the ability to create sound and complete provers, is not warranted. Even for decidable languages, the algorithmic complexity of provers may prohibit their exploitation.

To solve this problem a trade-off between the expressivity of the language and the complexity of its provers has to be found. These considerations have led to the definition of languages with limited complexity – like conceptual graphs and object-based representations – or of modular families of languages with associated modular prover algorithms – like description logics.

EXMO mainly considers languages with well-defined semantics (such as RDF and OWL that we contributed to define), and defines the semantics of some languages such as multimedia specification languages and alignment languages, in order to establish the properties of computer manipulations of the representations.

3.2. Ontology alignments

When different representations are used, it is necessary to identify their correspondences. This task is called ontology matching and its result is an alignment. It can be described as follows: given two ontologies, each describing a set of discrete entities (which can be classes, properties, rules, predicates, etc.), find the relationships, e.g., equivalence or subsumption, if any, that hold between these entities.

An alignment between two ontologies o and o' is a set of correspondences $\langle e, e', r \rangle$ in which:

- e and e' are the entities between which a relation is asserted by the correspondence, e.g., formulas, terms, classes, individuals;
- r is the relation asserted to hold between e and e'. This relation can be any relation applying to these entities, e.g., equivalence, subsumption.

In addition, a correspondence may support various types of metadata, in particular measures of the confidence in a correspodence.

Given the semantics of the two ontologies provided by their consequence relation, we define an interpretation of two aligned ontologies as a pair of interpretations $\langle m, m' \rangle$, one for each ontology. Such a pair of interpretations is a model of the aligned ontologies o and o' if and only if each respective interpretation is a model of the ontology and they satisfy all correspondences of the alignment.

This definition is extended to networks of ontologies: a set of ontologies and associated alignments. A model of such an ontology network is a tuple of local models such that each alignment is valid for the models involved in the tuple. In such a system, alignments play the role of model filters which will select the local models which are compatible with all alignments.

So, given an ontology network, it is possible to interpret it. However, given a set of ontologies, it is necessary to find the alignments between them and the semantics does not tell which ones they are. Ontology matching aims at finding these alignments. A variety of methods is used for this task. They perform pairwise comparisons of entities from each of the ontologies and select the most similar pairs. Most matching algorithms provide correspondences between named entities, more rarely between compound terms. The relationships are generally equivalence between these entities. Some systems are able to provide subsumption relations as well as other relations in the support language (like incompatibility or instanciation). Confidence measures are usually given a value between 0 and 1 and are used for expressing preferences between two correspondences.

GRAPHIK Project-Team

3. Scientific Foundations

3.1. Logic-based Knowledge Representation and Reasoning

We follow the mainstream *logical* approach to the KRR domain. First-order logic (FOL) is the reference logic in KRR and most formalisms in this area can be translated into fragments (i.e., particular subsets) of FOL. 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 KRR 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 *query answering* problem is a topical problem (see Sect. 3.3). It asks for the set of answers to a query in the KB. In the special 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 KRR 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 specific approach to KRR, which is based on labelled *graphs*. Indeed, we view labelled graphs as an *abstract representation* of knowledge that can be expressed in many KRR 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 reasonings 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. Ontological Query Answering

Querying knowledge bases is a central problem in knowledge representation and in database theory. 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. Classical ontogical languages, typically description logics, were not designed for efficient querying. On the other hand, database languages were 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 KRR:

- *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 some rules, both in their logical and their graph form, which are syntactically very simple but also very expressive. These rules can be seen as an abstraction of ontological knowledge expressed in the main languages used in the context of KB querying. See Sect. 6.1 for details on the results obtained.

A problem generalising the above described problems, and particularly relevant in the context of multiple data/metadata sources, is *querying hybrid knowledge bases*. In a hybrid knowledge base, each component may have its own formalism and its own reasoning mechanisms. There may be a common ontology shared by all components, or each component may have its own ontology, with mappings being defined among the ontologies. The question is what kind of interactions between these components and/or what limitations on the languages preserve the decidability of basic problems and if so, a "reasonable"complexity. Note that there are strong connections with data integration in databases.

3.4. Imperfect Information and Priorities

While classical FOL is the kernel of many KRR 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 in the agronomy domain (see Sect. 4.2):

1. to cope with vague and uncertain information and preferences in queries;
2. to cope with multi-granularity knowledge;
3. to take into account different and potentially conflicting viewpoints ;
4. to integrate decision notions (priorities, gravity, risk, benefit);
5. to integrate argumentation-based reasoning.

Although the solutions we will develop need 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. To achieve that goal, our main research directions are: non-monotonic reasonings (see ANR project ASPIQ in Sect. 8.1), as well as argumentation and preferences (see Sect. 6.2).

MAIA Project-Team

3. Scientific Foundations

3.1. Sequential Decision Making

3.1.1. Synopsis and Research Activities

Sequential decision making consists, in a nutshell, in controlling the actions of an agent facing a problem whose solution requires not one but a whole sequence of decisions. This kind of problem occurs in a multitude of forms. For example, important applications addressed in our work include: Robotics, where the agent is a physical entity moving in the real world; Medicine, where the agent can be an analytic device recommending tests and/or treatments; Computer Security, where the agent can be a virtual attacker trying to identify security holes in a given network; and Business Process Management, where the agent can provide an auto-completion facility helping to decide which steps to include into a new or revised process. Our work on such problems is characterized by three main research trends:

(A) *Understanding how, and to what extent, to best model the problems.*

(B) *Developing algorithms solving the problems and understanding their behavior.*

(C) *Applying our results to complex applications.*

Before we describe some details of our work, it is instructive to understand the basic forms of problems we are addressing. We characterize problems along the following main dimensions:

(1) Extent of the model: full vs. partial vs. none. This dimension concerns how complete we require the model of the problem – if any – to be. If the model is incomplete, then learning techniques are needed along with the decision making process.

(2) Form of the model: factored vs. enumerative. Enumerative models explicitly list all possible world states and the associated actions etc. Factored models can be exponentially more compact, describing states and actions in terms of their behavior with respect to a set of higher-level variables.

(3) World dynamics: deterministic vs. stochastic. This concerns our initial knowledge of the world the agent is acting in, as well as the dynamics of actions: is the outcome known a priori or are several outcomes possible?

(4) Observability: full vs. partial. This concerns our ability to observe what our actions actually do to the world, i.e., to observe properties of the new world state. Obviously, this is an issue only if the world dynamics are stochastic.

These dimensions are wide-spread in the AI literature. We remark that they are not exhaustive. In parts of our work, we also consider the difference between discrete/continuous problems, and centralized/decentralized problems. The complexity of solving the problem – both in theory and in practice – depends crucially on where the problem resides in this categorization. In many applications, not one but several points in the categorization make sense: simplified versions of the problem can be solved much more effectively and thus serve for the generation of *some* – if possibly sub-optimal – action strategy in a more feasible manner. Of course, the application as such may also come in different facets.

In what follows, we outline the main formal frameworks on which our work is based; while doing so, we highlight in a little more detail our core research questions. We then give a brief summary of how our work fits into the global research context.

3.1.2. Formal Frameworks

3.1.2.1. Deterministic Sequential Decision Making

Sequential decision making with deterministic world dynamics is most commonly known as **planning**, or **classical planning** [51]. Obviously, in such a setting every world state needs to be considered at most once, and thus enumerative models do not make sense (the problem description would have the same size as the space of possibilities to be explored). Planning approaches support factored description languages allowing to model complex problems in a compact way. Approaches to automatically learn such factored models do exist, however most works – and also most of our works on this form of sequential decision making – assume that the model is provided by the user of the planning technology. Formally, a problem instance, commonly referred to as a **planning task**, is a four-tuple $\langle V, A, I, G\rangle$. Here, V is a set of variables; a value assignment to the variables is a world state. A is a set of actions described in terms of two formulas over V: their preconditions and effects. I is the initial state, and G is a goal condition (again a formula over V). A solution, commonly referred to as a **plan**, is a schedule of actions that is applicable to I and achieves G.

Planning is **PSPACE**-complete even under strong restrictions on the formulas allowed in the planning task description. Research thus revolves around the development and understanding of search methods, which explore, in a variety of different ways, the space of possible action schedules. A particularly successful approach is **heuristic search**, where search is guided by information obtained in an automatically designed **relaxation** (simplified version) of the task. We investigate the design of relaxations, the connections between such design and the search space topology, and the construction of effective **planning systems** that exhibit good practical performance across a wide range of different inputs. Other important research lines concern the application of ideas successful in planning to stochastic sequential decision making (see next), and the development of technology supporting the user in model design.

3.1.2.2. Stochastic Sequential Decision Making

Markov Decision Processes (**MDP**) [52] are a natural framework for stochastic sequential decision making. An MDP is a four-tuple $\langle S, A, T, r\rangle$, where S is a set of states, A is a set of actions, $T(s, a, s') = P(s'|s, a)$ is the probability of transitioning to s' given that action a was chosen in state s, and $r(s, a, s')$ is the (possibly stochastic) reward obtained from taking action a in state s, and transitioning to state s'. In this framework, one looks for a **strategy**: a precise way for specifying the sequence of actions that induces, on average, an optimal sum of discounted rewards $\mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$. Here, $(r_0, r_1, ...)$ is the infinitely-long (random) sequence of rewards induced by the strategy, and $\gamma \in (0, 1)$ is a discount factor putting more weight on rewards obtained earlier. Central to the MDP framework is the Bellman equation, which characterizes the **optimal value function** V^*:

$$\forall s \in S, \quad V^*(s) = \max_{a \in A} \sum_{s' \in S} T(s, a, s')[r(s, a, s') + \gamma V^*(s')].$$

Once the optimal value function is computed, it is straightforward to derive an optimal strategy, which is deterministic and memoryless, i.e., a simple mapping from states to actions. Such a strategy is usually called a **policy**. An **optimal policy** is any policy π^* that is **greedy** with respect to V^*, i.e., which satisfies:

$$\forall s \in S, \quad \pi(s) \in \arg\max_{a \in A} \sum_{s' \in S} T(s, a, s')[r(s, a, s') + \gamma V^*(s')].$$

An important extension of MDPs, known as Partially Observable MDPs (**POMDPs**) allows to account for the fact that the state may not be fully available to the decision maker. While the goal is the same as in an MDP (optimizing the expected sum of discounted rewards), the solution is more intricate. Any POMDP can be seen to be equivalent to an MDP defined on the space of probability distributions on states, called **belief states**. The Bellman-machinery then applies to the belief states. The specific structure of the resulting MDP makes it possible to iteratively approximate the optimal value function – which is convex in the **belief space** – by piecewise linear functions, and to deduce an optimal policy that maps belief states to actions. A further

extension, known as a DEC-POMDP, considers $n \geq 2$ agents that need to control the state dynamics in a decentralized way without direct communication.

The MDP model described above is enumerative, and the complexity of computing the optimal value function is **polynomial** in the size of that input. However, in examples of practical size, that complexity is still too high so naïve approaches do not scale. We consider the following situations: (i) when the state space is large, we study approximation techniques from both a theoretical and practical point of view; (ii) when the model is unknown, we study how to learn an optimal policy from samples (this problem is also known as Reinforcement Learning [57]); (iii) in factored models, where MDP models are a strict generalization of classical planning – and are thus at least **PSPACE**-hard to solve – we consider using search heuristics adapted from such (classical) planning.

Solving a POMDP is **PSPACE**-hard even given an enumerative model. In this framework, we are mainly looking for assumptions that could be exploited to reduce the complexity of the problem at hand, for instance when some actions have no effect on the state dynamics (**active sensing**). The decentralized version, DEC-POMDPs, induces a significant increase in complexity (**NEXP**-complete). We tackle the challenging – even for (very) small state spaces – exact computation of finite-horizon optimal solutions through alternative reformulations of the problem. We also aim at proposing advanced heuristics to efficiently address problems with more agents and a longer time horizon.

3.1.3. Project-team positioning

Within Inria, the most closely related teams are TAO and Sequel. TAO works on evolutionary computation (EC) and statistical machine learning (ML), and their combination. Sequel works on ML, with a theoretical focus combining CS and applied maths. The main difference is that TAO and Sequel consider particular algorithmic frameworks that can, amongst others, be applied to Planning and Reinforcement Learning, whereas we revolve around Planning and Reinforcement Learning as the core problems to be tackled, with whichever framework suitable.

In France, we have recently begun collaborating with the IMS Team of Supélec Metz, notably with O. Pietquin and M. Geist who have a great expertise in approximate techniques for MDPs. We have have links with the MAD team of the BIA unit of the INRA at Toulouse, led by R. Sabbadin. They also use MDP related models and are interested in solving large size problems, but they are more driven by applications (mostly agricultural) than we are. In Paris, the Animat Lab, that was a part of the LIP6 and is now attached to the ISIR, has done some interesting works on factored Markov Decision Problems and POMDPs. Like us, their main goal was to tackle problems with large state space.

In Europe, the IDSIA Lab at Lugano (Switzerland) has brought some interesting ideas to the field of MDP (meta-learning, subgoal discovery) but seems now more interested in a *Universal Learner*. In Osnabrück (Germany), the Neuroinformatic group works on efficient reinforcement learning with a specific interest in the application to robotics. For deterministic planning, the most closely related groups are located in Freiburg (Germany), Glasgow (UK), and Barcelona (Spain). We have active collaborations with all of these.

In the rest of the world, the most important groups regarding MDPs can be found at Brown University, Rutgers Univ. (M. Littman), Univ. of Toronto (C. Boutilier), MIT AI Lab (L. Kaelbling, D. Bertsekas, J.Tsitsiklis), Stanford Univ., CMU, Univ. of Alberta (R. Sutton), Univ. of Massachusetts at Amherst (S. Zilberstein, A. Barto), *etc*. A major part of their work is aimed at making Markov Decision Process based tools work on real life problems and, as such, our scientific concerns meet theirs. For deterministic planning, important related groups and collaborators are to be found at NICTA (Canberra, Australia) and at Cornell University (USA).

3.2. Understanding and mastering complex systems

3.2.1. General context

There exist numerous examples of natural and artificial systems where self-organization and emergence occur. Such systems are composed of a set of simple entities interacting in a shared environment and exhibit complex collective behaviors resulting from the interactions of the local (or individual) behaviors of these entities.

The properties that they exhibit, for instance robustness, explain why their study has been growing, both in the academic and the industrial field. They are found in a wide panel of fields such as sociology (opinion dynamics in social networks), ecology (population dynamics), economy (financial markets, consumer behaviors), ethology (swarm intelligence, collective motion), cellular biology (cells/organ), computer networks (ad-hoc or P2P networks), etc.

More precisely, the systems we are interested in are characterized by :

- *locality*: Elementary components have only a partial perception of the system's state, similarly, a component can only modify its surrounding environment.
- *individual simplicity*: components have a simple behavior, in most cases it can be modeled by stimulus/response laws or by look-up tables. One way to estimate this simplicity is to count the number of stimulus/response rules for instance.
- *emergence*: It is generally difficult to predict the global behavior of the system from the local individual behaviors. This difficulty of prediction is often observed empirically and in some cases (e.g., cellular automata) one can show that the prediction of the global properties of a system is an undecidable problem. However, observations coming from simulations of the system may help us to find the regularities that occur in the system's behavior (even in a probabilistic meaning). Our interest is to work on problems where a full mathematical analysis seems out of reach and where it is useful to observe the system with large simulations. In return, it is frequent that the properties observed empirically are then studied on an analytical basis. This approach should allow us to understand more clearly where lies the frontier between simulation and analysis.
- *levels of description and observation*: Describing a complex system involves at least two levels: the micro level that regards how a component behaves, and the macro level associated with the collective behavior. Usually, understanding a complex system requires to link the description of a component behavior with the observation of a collective phenomenon: establishing this link may require various levels, which can be obtained only with a careful analysis of the system.

We now describe the type of models that are studied in our group.

3.2.2. Multi-agent models

To represent these complex systems, we made the choice to use reactive multi-agent systems (RMAS). Multi-agent systems are defined by a set of reactive agents, an environment, a set of interactions between agents and a resulting organization. They are characterized by a decentralized control shared among agents: each agent has an internal state, has access to local observations and influences the system through stimulus response rules. Thus, the collective behavior results from individual simplicity and successive actions and interactions of agents through the environment.

Reactive multi-agent systems present several advantages for modeling complex systems

- agents are explicitly represented in the system and have the properties of local action, interaction and observation;
- each agent can be described regardless of the description of the other agents, multi-agent systems allow explicit heterogeneity among agents which is often at the root of collective emergent phenomena;
- Multi-agent systems can be executed through simulation and provide good model to investigate the complex link between global and local phenomena for which analytic studies are hard to perform.

By proposing two different levels of description, the local level of the agents and the global level of the phenomenon, and several execution models, multi-agent systems constitute an interesting tool to study the link between local and global properties.

Despite of a widespread use of multi-agent systems, their framework still needs many improvements to be fully accessible to computer scientists from various backgrounds. For instance, there is no generic model to mathematically define a reactive multi-agent system and to describe its interactions. This situation is in contrast with the field of cellular automata, for instance, and underlines that a unification of multi-agent systems under a general framework is a question that still remains to be tackled. We now list the different challenges that, in part, contribute to such an objective.

3.2.3. Current challenges

Our work is structured around the following challenges that combine both theoretical and experimental approaches.

3.2.3.1. Providing formal frameworks

Currently, there is no agreement on a formal definition of a multi-agent system. Our research aims at translating the concepts from the field of complex systems into the multi-agent systems framework.

One objective of this research is to remove the potential ambiguities that can appear if one describes a system without explicitly formulating each aspect of the simulation framework. As a benefit, the reproduction of experiments is facilitated. Moreover, this approach is intended to gain a better insight of the self-organization properties of the systems.

Another important question consists in monitoring the evolution of complex systems. Our objective is to provide some quantitative characteristics of the system such as local or global stability, robustness, complexity, etc. Describing our models as dynamical systems leads us to use specific tools of this mathematical theory as well as statistical tools.

3.2.3.2. Controlling complex dynamical system

Since there is no central control of our systems, one question of interest is to know under which conditions it is possible to guarantee a given property when the system is subject to perturbations. We tackle this issue by designing exogenous control architectures where control actions are envisaged as perturbations in the system. As a consequence, we seek to develop control mechanism that can change the global behavior of a system without modifying the agent behavior (and not violating the autonomy property).

3.2.3.3. Designing systems

The aim is to design individual behaviors and interactions in order to produce a desired collective output. This output can be a collective pattern to reproduce in case of simulation of natural systems. In that case, from individual behaviors and interactions we study if (and how) the collective pattern is produced. We also tackle "inverse problems" (decentralized gathering problem, density classification problem, etc.) which consist in finding individual behaviors in order to solve a given problem.

3.2.4. Project-team positioning

Building a reactive multi-agent system consists in defining a set (generally a large number) of simple and reactive agents within a shared environment (physical or virtual) in which they move, act and interact with each other. Our interest in these systems is that, in spite of their simple definition at the agent level, they produce coherent and coordinated behavior at a global scale. The properties that they may exhibit, such as robustness and adaptivity explain why their study has been growing in the last decade (in the broader context of "complex systems").

Our work on such problems is characterized by five research trends: (A) *Defining a formal framework for describing and studying these systems*, (B) *Developing and understanding reactive multi-agent systems*, (C) *Analysing and proving properties*, (D) *Deploying these systems on typical distributed architectures such as swarms of robots, FPGAs, GPUs and sensor networks*, (E) *Transferring our results in applications.*

Multi-agent System is an active area of research in Artificial Intelligence and Complex Systems. Our research fits well into the international research context, and we have made and are making a variety of significant contributions both in theoretical and practical issues. Concerning multi-agent simulation and formalization, we compete or collaborate in France with S. Hassas in LIESP (Lyon), CERV (Brest), IREMIA (la Réunion), Ibisc (Evry), Lirmm (Montpellier), Irit (Toulouse), A. Drogoul (IRD, Bondy) and abroad with F. Zambonelli (Univ. Modena, Italy) A. Deutsch (Dresden, Germany), D. Van Parunak (Vector research, USA), P. Valkenaers, D. Weyns (Univ. Leuven, Belgium), etc. Regarding our work on swarm robotics we have common objectives with the DISAL [3] EPFL Laboratory, the Bristol Robotics Laboratory, the Distributed Robotics Laboratory at MIT, the team of W. & D. Spears at Wyoming university, the Pheromone Robotics project at HRL Lab. [4], the FlockBots project at GMU[5], the team of G. Théraulaz at CNRS-Toulouse and the teams of J.-L. Deneubourg and M. Dorigo at ULB (Bruxelles).

[3] Distributed Intelligent Systems and Algorithms Laboratory including EPFL Swarm-Intelligent Systems Group (SWIS) founded in 2003 and the Collective Robotics Group (CORO) founded in 2000 at California Institute of Technology USA

[4] HRL, Information and systems sciences Lab (ISSL), Malibu CA, USA (D. Payton)

[5] George Mason University, Eclab, USA (L. Panait, S. Luke)

MOSTRARE Project-Team

3. Scientific Foundations

3.1. Modeling XML document transformations

Participants: Guillaume Bagan, Adrien Boiret, Iovka Boneva, Angela Bonifati, Anne-Cécile Caron, Benoît Groz, Joachim Niehren, Yves Roos, Sławek Staworko, Sophie Tison, Antoine Ndione, Tom Sebastian.

XML document transformations can be defined in W3C standards languages XQuery or XSLT. Programming XML transformations in these languages is often difficult and error prone even if the schemata of input and output documents are known. Advanced programming experience and considerable programming time may be necessary, that are not available in Web services or similar scenarios.

Alternative programming language for defining XML transformations have been proposed by the programming language community, for instance XDuce [46], Xtatic [44], [49], and CDuce [35], [36], [37]. The type systems of these languages simplify the programming tasks considerably. But of course, they don't solve the general difficulty in programming XML transformations manually.

Languages for defining node selection queries arise as sub-language of all XML transformation languages. The W3C standards use XPath for defining monadic queries, while XDuce and CDuce rely on regular queries defined by regular pattern equivalent to tree automata. Indeed, it is natural to look at node selection as a simple form of tree transformation. Monadic node selection queries correspond to deterministic transformations that annotate all selected nodes positively and all others negatively. N-ary node selection queries become non-deterministic transformations, yielding trees annotated by Boolean vectors.

After extensive studies of node selection queries in trees (in XPath and many other languages) the XML community has started more recently to formally investigate XML tree transformations. The expressiveness and complexity of XQuery are studied in [48], [57]. Type preservation is another problem, i.e., whether all trees of the input type get transformed into the output type, or vice versa, whether the inverse image of the output type is contained in the input type [52], [50].

The automata community usually approaches tree transformations by tree transducers [42], i.e., tree automata producing output structure. Macro tree transducers, for instance, have been proposed recently for defining XML transformations [50]. From the view point of logic, tree transducers have been studied for MSO definability [43].

3.2. Machine learning for XML document transformations

Participants: Jean Decoster, Pascal Denis, Jean-Baptiste Faddoul, Rémi Gilleron, Grégoire Laurence, Aurélien Lemay, Joachim Niehren, Sławek Staworko, Marc Tommasi, Fabien Torre, Gemma Garriga, Antonino Freno, Thomas Ricatte, Mikaela Keller.

Automatic or semi-automatic tools for inferring tree transformations are needed for information extraction. Annotated examples may support the learning process. The learning target will be models of XML tree transformations specified in some of the languages discussed above.

Grammatical inference is commonly used to learn languages from examples and can be applied to learn transductions. Previous work on grammatical inference for transducers remains limited to the case of strings [38], [53]. For the tree case, so far only very basic tree transducers have been shown to be learnable, by previous work of the Mostrare project. These are node selecting tree transducer (NSTTs) which preserve the structure of trees while relabeling their nodes deterministically.

Statistical inference is most appropriate for dealing with uncertain or noisy data. It is generally useful for information extraction from textual data given that current text understanding tools are still very much limited. XML transformations with noisy input data typically arise in data integration tasks, as for instance when converting PDF into XML.

Stochastic tree transducers have been studied in the context of natural language processing [45], [47]. A set of pairs of input and output trees defines a relation that can be represented by a 2-tape automaton called a *stochastic finite-state transducer* (SFST). A major problem consists in estimating the parameters of such transducer. SFST training algorithms are lacking so far [41].

Probabilistic context free grammars (pCFGs) [51] are used in the context of PDF to XML conversion [39]. In the first step, a labeling procedure of leaves of the input document by labels of the output DTD is learned. In the second step, given a CFG as a generative model of output documents, probabilities are learned. Such two steps approaches are in competition with one step approaches estimating conditional probabilities directly.

A popular non generative model for information extraction is *conditional random fields* (CRF, see a survey [54]). One main advantage of CRF is to take into account long distance dependencies in the observed data. CRF have been defined for general graphs but have mainly been applied to sequences, thus CRF for XML trees should be investigated.

So called *structured output* has recently become a research topic in machine learning [56], [55]. It aims at extending the classical categorization task, which consists to associate one or some labels to each input example, in order to handle structured output labels such as trees. Applicability of structured output learning algorithms remains to be asserted for real tasks such as XML transformations.

OAK Team

3. Scientific Foundations

3.1. Efficient XML and RDF data management

The development of Web technologies has led to a strong increase in the number and complexity of the applications which represent their data in Web formats, among which XML (for structured documents) and RDF (for Semantic Web data) are the most prominent. Oak has carried on research on algorithms and systems for efficiently processing expressive queries on such Web data formats. We have considered the efficient management of XML and RDF data, both for query evaluation and for efficiently applying updates, possibly in concurrence with queries. We have also started investigating multidimensional data analysis within RDF data warehouses.

For applications that integrate such Web data from various sources, we developed efficient and effective techniques to automatically recognize multiple representations of the same real-world object. That is, we devised main-memory resident algorithms that apply on hierarchical data [8] as well as algorithms that manipulate graph data leveraging off-the-shelf database management systems and parallelization to address both efficiency and scalability beyond main-memory [7].

3.2. Cloud-based Data Management

We have recently started to work on the efficient management of complex Web data, in particular structured XML documents and Semantic Web data under the form of RDF, in a cloud-based data management platform. We have investigated architectures and algorithms for storing Web data in elastic cloud-based stores and building an index for such data within the efficient key-value stores provided by off-the-shelf cloud data management platform. We have devised and prototyped such platforms for both XML and RDF data, and started experimenting with them in the Amazon Web Services (AWS) platform [13], [12], [10].

3.3. Data Transformation Management

With the increasing complexity of data processing queries, for instance in applications such as relational data analysis or integration of Web data (e.g., XML or RDF) comes the need to better manage complex data transformations. This includes systematically verifying, maintaining, and testing the transformations an application relies on. In this context, Oak has focused on verifying the semantic correctness of a declarative program that specifies a data transformation query, e.g., an SQL query. To this end, we have investigated how to leverage data provenance (the information of the origin of data and the query operators) for query debugging. More specifically, we developed and implemented algorithms to explain unexpected results produced by a query (why-provenance) as well as expected results that are however missing from the query result (why-not provenance). Results have been presented in form of a software prototype [15].

ORPAILLEUR Project-Team

3. Scientific Foundations

3.1. From KDD to KDDK

knowledge discovery in databases, knowledge discovery in databases guided by domain knowledge, data mining methods

Knowledge discovery in databases is a process for extracting knowledge units from large databases, units that can be interpreted and reused within knowledge-based systems.

Knowledge discovery in databases is a process for extracting knowledge units from large databases, units that can be interpreted and reused within knowledge-based systems. From an operational point of view, the KDD process is performed within a KDD system including databases, data mining modules, and interfaces for interactions, e.g. editing and visualization. The KDD process is based on three main operations: selection and preparation of the data, data mining, and finally interpretation of the extracted units. The KDDK process –as implemented in the research work of the Orpailleur team– is based on *data mining methods* that are either symbolic or numerical:

- Symbolic methods are based on frequent itemsets search, association rule extraction, and concept lattice design (Formal Concept Analysis and extensions [95], [108]).
- Numerical methods are based on second-order Hidden Markov Models (HMM2, designed for pattern recognition [107]). Hidden Markov Models have good capabilities for locating stationary segments, and are mainly used for mining temporal and spatial data.

The principle summarizing KDDK can be understood as a process going from complex data units to knowledge units being guided by domain knowledge (KDDK or "knowledge with/for knowledge") [104]. Two original aspects can be underlined: (i) the KDD process is guided by domain knowledge, and (ii) the extracted units are embedded within a knowledge representation formalism to be reused in a knowledge-based system for problem solving purposes.

The various instantiations of the KDDK process in the research work of Orpailleur are mainly based on *classification*, considered as a polymorphic process involved in tasks such as modeling, mining, representing, and reasoning. Accordingly, the KDDK process may feed knowledge-based systems to be used for problem-solving activities in application domains, e.g. agronomy, astronomy, biology, chemistry, and medicine, and also for semantic web activities involving text mining, information retrieval, and ontology engineering [97], [81].

3.2. Methods for Knowledge Discovery guided by Domain Knowledge

knowledge discovery in databases guided by domain knowledge, lattice-based classification, formal concept analysis, frequent itemset search, association rule extraction, second-order Hidden Markov Models, stochastic process, numerical data mining method

knowledge discovery in databases guided by domain knowledge is a KDD process guided by domain knowledge ; the extracted units are represented within a knowledge representation formalism and embedded within a knowledge-based system.

Classification problems can be formalized by means of a class of individuals (or objects), a class of properties (or attributes), and a binary correspondence between the two classes, indicating for each individual-property pair whether the property applies to the individual or not. The properties may be features that are present or absent, or the values of a property that have been transformed into binary variables. Formal Concept Analysis (FCA) relies on the analysis of such binary tables and may be considered as a symbolic data mining technique to be used for extracting (from a binary database) a set of formal concepts organized within a concept lattice [95]. Concept lattices are sometimes also called Galois lattices [82].

The search for frequent itemsets and the extraction of association rules are well-known symbolic data mining methods, related to FCA (actually searching for frequent itemsets may be understood as traversing a concept lattice). Both processes usually produce a large number of items and rules, leading to the associated problems of "mining the sets of extracted items and rules". Some subsets of itemsets, e.g. frequent closed itemsets (FCIs), allow to find interesting subsets of association rules, e.g. informative association rules. This is why several algorithms are needed for mining data depending on specific applications [123], [122].

Among useful patterns extracted from a database, frequent itemsets are usually thought to unfold "regularities" in the data, i.e. they are the witnesses of recurrent phenomena and they are consistent with the expectations of the domain experts. In some situations however, it may be interesting to search for "rare" itemsets, i.e. itemsets that do not occur frequently in the data (contrasting frequent itemsets). These correspond to unexpected phenomena, possibly contradicting beliefs in the domain. In this way, rare itemsets are related to "exceptions" and thus may convey information of high interest for experts in domains such as biology or medicine [124], [125].

From the numerical point of view, a Hidden Markov Model (HMM2) is a stochastic process aimed at extracting and modeling a sequence of stationary distributions of events. These models can be used for data mining purposes, especially for spatial and temporal data as they show good capabilities to locate patterns both in time and space domains. A special research effort focuses on the combination of knowledge elicited by experts and time-space regularities as extracted by an unsupervised classification based on stochastic models [23].

3.3. Elements on Text Mining

knowledge discovery form large collection of texts, text mining, information extraction, document annotation, ontologies

Text mining is a process for extracting knowledge units from large collections of texts, units that can be interpreted and reused within knowledge-based systems.

The objective of a text mining process is to extract new and useful knowledge units in a large set of texts [80], [89]. 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 the mining process more complex. To avoid information dispersion, 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 KDDK perspective, the text mining process is aimed at extracting new knowledge units from texts with the help of background knowledge encoded within an ontology and which is useful to relate notions present in a text, to guide and to help the text mining process. Text mining is especially useful in the context of semantic web for ontology engineering [86], [85], [84]. In the Orpailleur team, the focus is put on real-world texts in application domains such as astronomy, biology and medicine, using mainly 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".

3.4. Elements on Knowledge Systems and Semantic Web

knowledge representation, ontology, description logics, classification-based reasoning, case-based reasoning, semantic web, knowledge-based information retrieval, web mining

Knowledge representation is a process for representing knowledge within an ontology using a knowledge representation formalism, giving knowledge units a syntax and a semantics. Semantic web is based on ontologies and allows search, manipulation, and dissemination of documents on the web by taking into account their contents, i.e. the semantics of the elements included in the documents.

Usually, people try to take advantage of the web by searching for information (navigation, exploration), and by querying documents using search engines (information retrieval). Then people try to analyze the obtained results, a task that may be very difficult and tedious. Semantic web is an attempt for guiding search for information with the help of machines, that are in charge of asking questions, searching for answers, classifying and interpreting the answers. However, a machine 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 ontologies are of main importance with respect to the task of setting up semantic web. Thus, there is a need for representation languages for annotating documents, i.e. describing the content of documents, and giving a semantics to this content. Knowledge representation languages are (the?) good candidates for achieving the task: they have a syntax with an associated semantics, and they can be used for retrieving information, answering queries, and reasoning.

Semantic web constitutes a good platform for experimenting ideas on knowledge representation, reasoning, and KDDK. In particular, the knowledge representation language used for designing ontologies is the OWL language, which is based on description logics (or DL [79]). In OWL, knowledge units are represented within concepts (or classes), with attributes (properties of concepts, or relations, or roles), and individuals. The hierarchical organization of concepts (and relations) relies on a subsumption relation that is a partial ordering. The inference services are based on subsumption, concept and individual classification, two tasks related to "classification-based reasoning". Furthermore, classification-based reasoning can be associated to case-based reasoning (CBR), that relies on three main operations: retrieval, adaptation, and memorization. Given a target problem, retrieval consists in searching for a source (memorized) problem similar to the target problem. Then, the solution of the source problem is adapted to fulfill the constraints attached to the target problem, and possibly memorized for further reuse.

In the trend of semantic web, research work is also carried on semantic wikis which are wikis i.e., web sites for collaborative editing, in which documents can be annotated thanks to semantic annotations and typed relations between wiki pages. Such links provide kind of primitive knowledge units that can be used for guiding information retrieval or knowledge discovery.

SMIS Project-Team

3. Scientific Foundations

3.1. Embedded Data Management

The challenge tackled is this research action is twofold: (1) to design embedded database techniques matching the hardware constraints of (current and future) smart objects and (2) to set up co-design rules helping hardware manufacturers to calibrate their future platforms to match the requirements of data driven applications. While a large body of work has been conducted on data management techniques for high-end servers (storage, indexation and query optimization models minimizing the I/O bottleneck, parallel DBMS, main memory DBMS, etc.), less research efforts have been placed on embedded database techniques. Light versions of popular DBMS have been designed for powerful handheld devices yet DBMS vendors have never addressed the complex problem of embedding database components into chips. Proposals dedicated to databases embedded on chip usually consider small databases, stored in the non-volatile memory of the microcontroller –hundreds of kilobytes– and rely on NOR Flash or EEPROM technologies. Conversely, SMIS is pioneering the combination of microcontrollers and NAND Flash constraints to manage Gigabyte(s) size embedded databases. We present below the positioning of SMIS with respect to international teams conducting research on topics which may be connected to the addressed problem, namely work on electronic stable storage, RAM consumption and specific hardware platforms.

Major database teams are investigating data management issues related to hardware advances (EPFL: A. Ailamaki, CWI: M. Kersten, U. Of Wisconsin: J. M. Patel, Columbia: K. Ross, UCSB: A. El Abbadi, IBM Almaden: C. Mohan, etc.). While there are obvious links with our research on embedded databases, these teams target high-end computers and do not consider highly constrained architectures with non traditional hardware resources balance. At the other extreme, sensors (ultra-light computing devices) are considered by several research teams (e.g., UC Berkeley: D. Culler, ITU: P. Bonnet, Johns Hopkins University: A. Terzis, MIT: S. Madden, etc.). The focus is on the processing of continuous streams of collected data. Although the devices we consider share some hardware constraints with sensors, the objectives of both environments strongly diverge in terms of data cardinality and complexity, query complexity and data confidentiality requirements. Several teams are looking at efficient indexes on flash (HP LABS: G. Graefe, U. Minnesota: B. Debnath, U. Massachusetts: Y. Diao, Microsoft: S. Nath, etc.). Some studies try to minimize the RAM consumption, but the considered RAM/stable storage ratio is quite large compared to the constraints of the embedded context. Finally, a large number of teams have focused on the impact of flash memory on database system design (we presented an exhaustive state of the art in a VLDB tutorial [7]). The work conducted in the SMIS team on bi-modal flash devices takes the opposite direction, proposing to influence the design of flash devices by the expression of database requirements instead of running after the constantly evolving flash device technology.

3.2. Access and Usage Control Models

Access control management has been deeply studied for decades. Different models have been proposed to declare and administer access control policies, like DAC, MAC, RBAC, TMAC, and OrBAC. While access control management is well established, new models are being defined to cope with privacy requirements. Privacy management distinguishes itself from traditional access control is the sense that the data to be protected is personal. Hence, the user's consent must be reflected in the access control policies, as well as the usage of the data, its collection rules and its retention period, which are principles safeguarded by law and must be controlled carefully.

The research community working on privacy models is broad, and involves many teams worldwide including in France ENST-B, LIRIS, Inria LICIT, and LRI, and at the international level IBM Almaden, Purdue Univ., Politecnico di Milano and Univ. of Milano, George Mason Univ., Univ. of Massachusetts, Univ. of Texas and Colorado State Univ. to cite a few. Pioneer attempts towards privacy wary systems include the P3P Platform for Privacy Preservation [39] and Hippocratic databases [30]. In the last years, many other policy languages have been proposed for different application scenarios, including EPAL [44], XACML [41] and WSPL [34]. Hippocratic databases are inspired by the axiom that databases should be responsible for the privacy preservation of the data they manage. The architecture of a Hippocratic database is based on ten guiding principles derived from privacy laws.

The trend worldwide has been to propose enhanced access control policies to capture finer behaviour and bridge the gap with privacy policies. To cite a few, Ardagna *et al.* (Univ. Milano) enables actions to be performed after data collection (like notification or removal), purpose binding features have been studied by Lefevre *et al.* (IBM Almaden), and Ni *et al.* (Purdue Univ.) have proposed obligations and have extended the widely used RBAC model to support privacy policies.

The positioning of the SMIS team within this broad area is rather (1) to focus on intuitive or automatic tools helping the individual to control some facets of her privacy (e.g., data retention, minimal collection) instead of increasing the expressiveness but also the complexity of privacy models and (2) to push concrete models enriched by real-case (e.g., medical) scenarios and by a joint work with researchers in Law.

3.3. Tamper-resistant Data Management

Tamper-resistance refers to the capacity of a system to defeat confidentiality and integrity attacks. This problem is complementary to access control management while being (mostly) orthogonal to the way access control policies are defined. Security surveys regularly point out the vulnerability of database servers against external (i.e., by intruders) and internal (i.e., by employees) attacks. Several attempts have been made in commercial DBMSs to strengthen server-based security, e.g., by separating the duty between DBA and DSA (Data Security Administrator), by encrypting the database footprint and by securing the cryptographic material using Hardware Security Modules (HSM) [36]. To face internal attacks, client-based security approaches have been investigated where the data is stored encrypted on the server and is decrypted only on the client side. Several contributions have been made in this direction, notably by U. of California Irvine (S. Mehrotra, Database Service Provider model), IBM Almaden (R. Agrawal, computation on encrypted data), U. of Milano (E. Damiani, encryption schemes), Purdue U. (E. Bertino, XML secure publication), U. of Washington (D. Suciu, provisional access) to cite a few seminal works. An alternative, recently promoted by Stony Brook Univ. (R. Sion), is to augment the security of the server by associating it with a tamper-resistant hardware module in charge of the security aspects. Contrary to traditional HSM, this module takes part in the query computation and performs all data decryption operations. SMIS investigates another direction based on the use of a tamper-resistant hardware module on the client side. Most of our contributions in this area are based on exploiting the tamper-resistance of secure tokens to build new data protection schemes.

While our work on Privacy-Preserving data Publishing (PPDP) is still related to tamper-resistance, a complementary positioning is required for this specific topic. The primary goal of PPDP is to anonymize/sanitize microdata sets before publishing them to serve statistical analysis purposes. PPDP (and privacy in databases in general) is a hot topic since 2000, when it was introduced by IBM Research (R. Agrawal : IBM Almaden, C.C. Aggarwal: IBM Watson), and many teams, mostly north American universities or research centres, study this topic (e.g., PORTIA DB-Privacy project regrouping universities such as Stanford with H. Garcia-Molina). Much effort has been devoted by the scientific community to the definition of privacy models exhibiting better privacy guarantees or better utility or a balance of both (such as differential privacy studied by C. Dwork : Microsoft Research or D. Kifer : Penn-State Univ and J. Gehrke : Cornell Univ) and thorough surveys exist that provide a large overview of existing PPDP models and mechanisms [40]. These works are however orthogonal to our approach in that they make the hypothesis of a trustworthy central server that can execute the anonymization process. In our work, this is not the case. We consider an architecture composed of a large

population of tamper-resistant devices weakly connected to an untrusted infrastructure and study how to compute PPDP problems in this context. Hence, our work has some connections with the works done on Privacy Preserving Data Collection (R.N.Wright : Stevens Institute of Tech. / Rutgers Univ, NJ, V. Shmatikov : Univ Austin Texas), on Secure Multi-party Computing for Privacy Preserving Data Mining (J. Vaidya : Rutgers Univ, C. Clifton : Purdue Univ) and on distributed PPDP algorithms (D. DeWitt : Univ Wisconsin, K. Lefevre : Univ Michigan, J. Vaidya : Rutgers Univ, C. Clifton : Purdue Univ) while none of them share the same architectural hypothesis as us.

WAM Project-Team

3. Scientific Foundations

3.1. XML Processing

Participants: Melisachew Chekol, Pierre Genevès, Nils Gesbert, Nicola Guido, Muhammad Junedi, Nabil Layaïda, Manh-Toan Nguyen, Vincent Quint.

Extensible Markup Language (XML) has gained considerable interest from industry, and plays now a central role in modern information system infrastructures. In particular, XML is the key technology for describing, storing, and exchanging a wide variety of data on the web. The essence of XML consists in organizing information in tree-tagged structures conforming to some constraints which are expressed using type languages such as DTDs, XML Schemas, and Relax NG.

There still exist important obstacles in XML programming, especially in the areas of performance and reliability. Programmers are given two options: domain-specific languages such as XSLT, or general-purpose languages augmented with XML application programming interfaces such as the Document Object Model (DOM). Neither of these options is a satisfactory answer to performance and reliability issues, nor is there even a trade-off between the two. As a consequence, new paradigms are being proposed which all have the aim of incorporating XML data as first-class constructs in programming languages. The hope is to build a new generation of tools that are capable of taking reliability and performance into account at compile time.

One of the major challenges in this line of research is to develop automated and tractable techniques for ensuring static type safety and optimization of programs. To this end, there is a need to solve some basic reasoning tasks that involve very complex constructions such as XML types (regular tree types) and powerful navigational primitives (XPath expressions or CSS selectors). In particular, every future compiler of XML programs will have to routinely solve problems such as:

- XPath query emptiness in the presence of a schema: if one can decide at compile time that a query is not satisfiable, then subsequent bound computations can be avoided
- query equivalence, which is important for query reformulation and optimization
- path type-checking, for ensuring at compile time that invalid documents can never arise as the output of XML processing code.

All these problems are known to be computationally heavy (when decidable), and the related algorithms are often tricky.

We have developed an XML/XPath static analyzer based on a new logic of finite trees. This analyzer consists of:

- compilers that allow XML types, XPath queries, and CSS selectors to be translated into this logic
- an optimized logical solver for testing satisfiability of a formula of this logic.

The benefit of these compilers is that they allow one to reduce all the problems listed above, and many others too, to logical satisfiability. This approach has a couple of important practical advantages. First of all, one can use the satisfiability algorithm to solve all of these problems. More importantly, one could easily explore new variants of these problems, generated for example by the presence of different kinds of type or schema information, with no need to devise a new algorithm for each variant.

3.2. Multimedia Models and Languages

Participants: Nicolas Hairon, Yohan Lasorsa, Nabil Layaïda, Jacques Lemordant, Vincent Quint, Cécile Roisin.

We have participated in the international endeavor for defining a standard multimedia document format for the web that accommodates the constraints of different types of terminals. SMIL is the main outcome of this work. It focuses on a modular and scalable XML format that combines efficiently the different dimensions of a multimedia web document: synchronization, layout and linking. Our current work on multimedia formats follows the same trend.

With the advent of HTML5 and its support in all popular browsers, HTML is becoming an important multimedia language. Video and audio can now be embedded in HTML pages without worrying about the availability of plugins. However, animation and synchronization of a HTML5 page still require programming skills. To address this issue, we are developing a scheduler that allows HTML documents to be animated and synchronized in a purely declarative way. This work is based on the SMIL Timing and Synchronization module and the SMIL Timesheets specification. The scheduler is implemented in JavaScript, which makes it usable in any browser. Timesheets can also be used with other XML document languages, such as SVG for instance.

Audio is the poor relation in the web format family. Most contents on the web may be represented in a structured way, such as text in HTML or XML, graphics in SVG, or mathematics in MathML, but sound was left aside with low-level representations that basically only encode the audio signal. Our work on audio formats aims at allowing sound to be on a par with other contents, in such a way it could be easily combined with them in rich multimedia documents that can then be processed safely in advanced applications. More specifically, we have participated in IAsig (Interactive Audio special interest group), an international initiative for creating a new format for interactive audio called iXMF (Interactive eXtensible Music Format). We are now developing A2ML, an XML format for embedded interactive audio, deriving from well-established formats such as iXMF and SMIL. We use it in augmented environments (see section 3.4), where virtual, interactive, 3D sounds are combined with the real sonic environment.

Regarding discrete media objects in multimedia documents, popular document languages such as HTML can represent a very broad range of documents, because they contain very general elements that can be used in many different situations. This advantage comes at the price of a low level of semantics attached to the structure. The concepts of microformats and semantic HTML were proposed to tackle this weakness. More recently, RDFa and microdata were introduced with the same goal. These formats add semantics to web pages while taking advantage of the existing HTML infrastructure. With this approach new applications can be deployed smoothly on the web, but authors of web pages have very little help for creating and encoding this kind of semantic markup. A language that addresses these issues is developed and implemented in WAM. Called XTiger, its role is to specify semantically rich XML languages in terms of other, less expressive XML languages, such as HTML. Recent extensions to the language make it now usable also to edit pure XML documents and to define their structure model (see section 3.3).

3.3. Multimedia Authoring

Participants: Nicolas Hairon, Yohan Lasorsa, Jacques Lemordant, David Liodenot, Vincent Quint, Mathieu Razafimahazo, Cécile Roisin.

3.3.1. Structured editing

Multimedia documents are considered through several kinds of structures: logical organization, layout, time, linking, animations. We are working on techniques that allow authors of such documents to manipulate all these structures in homogeneous environments. The main objective is to support new advances in document formats without making the authoring task more complex. The key idea is to present simultaneously several views of the document, each view putting the emphasis on a particular structure, and to allow authors to manipulate each view directly and efficiently. As the various structures of a document are not independent from each other, views are “synchronized” to reflect in each of them the consequences of every change made in a particular view. The XML markup, although it can be accessed at any time, is handled by the tools, and authors do not have to worry about syntactical issues.

3.3.2. Template-driven editing

We have more recently experimented another way to edit highly structured XML documents without the usual complexity of the most common XML editors. The novelty of the approach is to use templates instead of XML schemas or DTDs, and to run the editor as a web application, within the browser. This way, it is much easier to create new document types and to provide an editing environment for these document types, that any web user can instantly use. This lightweight approach to XML editing complements the previous approach by covering new categories of XML applications.

3.4. Augmented Environments

Participants: Yohan Lasorsa, Jacques Lemordant, David Liodenot, Thibaud Michel, Mathieu Razafimahazo.

The term Augmented Environments refers collectively to ubiquitous computing, context-aware computing, and intelligent environments. The goal of our research on these environments is to introduce personal Augmented Reality (AR) devices, taking advantage of their embedded sensors. We believe that personal AR devices such as mobile phones or tablets will play a central role in augmented environments. These environments offer the possibility of using ubiquitous computation, communication, and sensing to present context-sensitive information and services to the user.

AR applications often rely on 3D content and employ specialized hardware and computer vision techniques for both tracking and scene reconstruction. Our approach tries to seek a balance between these traditional AR contexts and what has come to be known as mobile AR browsing. It first acknowledges that mobile augmented environment browsing does not require that 3D content be the primary means of authoring. It provides instead a method for HTML5 and audio content to be authored, positioned in the surrounding environments and manipulated as freely as in modern web browsers.

Many service providers of augmented environments desire to create innovative services. Accessibility of buildings is one example we are involved in. However, service providers often have to strongly rely on experience, intuition, and tacit knowledge due to lack of tools on which to base a scientific approach. Augmented environments offer the required rigorous approach that enables Evidence-Based Services (EBS) if adequate tools for AR technologies are designed. Service cooperation through exchange of normalized real-time data or data logs is one of these tools, together with sensor data streams fusion inside an AR mobile browser. EBS can improve the performance of real-world sensing, and conversely EBS models authoring and service operation can be facilitated by real-world sensing.

The applications we use to elaborate and validate our concepts are pedestrian navigation for visually impaired people and applications for cultural heritage visits. On the authoring side, we are interested in interactive indoor modeling, audio mobile mixing, and formats for Points of Interest. Augmented environment services we consider are, among others, behavior analysis for accessibility, location services, and indoor geographical information services.

WIMMICS Team

3. Scientific Foundations

3.1. Analyzing and Modeling Users, Communities and their Interactions in a Social Semantic Web Context

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.

In the short term, following the user modeling technique known as *Personas*, we are interested in these user models that 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 will specialize *Personas* approaches to include aspects appropriate to Web applications. The formalization of these models will rely on ontology-based modeling of users and communities starting with generalist schemas (e.g. FOAF: *Friend of a Friend*). In the longer term we will consider additional extensions of these schemas to capture additional aspects (e.g. emotional states). We will extend current descriptions of relational and emotional aspects in existing variants of the *Personas* technique.

Beyond the individual user models, we propose to 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. In the short term we will further develop our method for elaborating collective personas and compare it to the related *collaboration personas* method and to the group modeling methods which are extensions to groups of the classical user modeling techniques dedicated to individuals. We also propose to rely on and adapt participatory sketching and prototyping to support the design of interfaces for visualizing and manipulating representations of collectives. In the longer term we want to focus on studying and modeling 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.

Since we have a background in requirement models, we want to consider in the short term their formalization too in order to support mutual understanding and interoperability between requirements expressed with these heterogeneous models. In a longer term, we believe that argumentation theory can be combined to requirement engineering to improve participant awareness and support decision-making. On the methodological side, we propose to adapt to the design of such systems the incremental formalization approach originally introduced in the context of CSCW (Computer Supported Cooperative Work) and HCI (Human Computer Interaction) communities.

Finally, in the short term, for all the models we identified here we will rely on and evaluate knowledge representation methodologies and theories, in particular ontology-based modeling. In the longer term, additional models of the contexts, devices, processes and mediums will also be formalized and used to support adaptation, proof and explanation and foster acceptation and trust from the users. We specifically target a unified formalization of these contextual aspects to be able to integrate them at any stage of the processing.

3.2. Formalizing and Reasoning on Heterogeneous Semantic Graphs

Our second line of work is to formalize as typed graphs the models identified in the previous section in order for software to exploit them in their processing. The challenge then is two-sided:

- To propose models and formalisms to capture and merge representations of both kinds 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.

- To propose algorithms (in particular graph-based reasoning) and approaches (e.g. human-computing methods) to process these mixed representations. In particular we are interested in allowing cross-enrichment between them and in exploiting the life cycle and specificities of each one to foster the life-cycles of the others.

While some of these problems are known, for instance in the field of knowledge representation and acquisition (e.g. disambiguation, fuzzy representations, argumentation theory), the Web reopens them with exacerbated difficulties of scale, speed, heterogeneity, and an open-world assumption.

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. Scientific Foundations

3.1. 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 DBMSs, now at the heart of any information system, and of advanced data management capabilities in many kinds of software products (application servers, document systems, search engines, directories, etc.).

The fundamental principle behind data management is *data independence*, which enables applications and users to deal with the data at a high conceptual level while ignoring implementation details. The relational model, by resting on a strong theory (set theory and first-order logic) to provide data independence, has revolutionized data management. The major innovation of relational DBMS has been to allow data manipulation through queries expressed in a high-level (declarative) language such as SQL. Queries can then be automatically translated into optimized query plans that take advantage of underlying access methods and indices. Many other advanced capabilities have been made possible by data independence : data and metadata modeling, schema management, consistency through integrity rules and triggers, transaction support, etc.

This data independence principle has also enabled DBMS to continuously integrate new advanced capabilities such as object and XML support and to adapt to all kinds of hardware/software platforms from very small smart devices (smart phone, PDA, smart card, etc.) to very large computers (multiprocessor, cluster, etc.) in distributed environments.

Following the invention of the relational model, research in data management has continued with the elaboration of strong database theory (query languages, schema normalization, complexity of data management algorithms, transaction theory, etc.) and the design and implementation of DBMS. For a long time, the focus was on providing advanced database capabilities with good performance, for both transaction processing and decision support applications. And the main objective was to support all these capabilities within a single DBMS.

The problems of scientific data management (massive scale, complexity and heterogeneity) go well beyond the traditional context of DBMS. To address them, we capitalize on scientific foundations in closely related domains: distributed data management, cloud data management, uncertain data management, metadata integration, data mining and content-based information retrieval.

3.2. Distributed Data Management

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, to address issues in data integration, scientific workflows, recommendation, query processing and data analysis.

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) [13]. Transparency is achieved through a global schema which hides the local databases' heterogeneity. In its simplest form, a distributed database system is a centralized server that supports a global schema and implements distributed database techniques (query processing, transaction management, consistency management, etc.). This approach has proved 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. Data integration systems, e.g. price comparators such as KelKoo, extend the distributed database approach to access data sources on the Internet with a simpler query language in read-only mode.

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.

Scientific workflow management systems (SWfMS) such as Kepler (http://kepler-project.org) and Taverna (http://www.taverna.org.uk) allow scientists to describe and execute complex scientific procedures and 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 and demand high performance computing (HPC) environments with highly distributed data sources and computing resources. However, combining SWfMS with HPC to improve throughput and performance remains a difficult challenge. In particular, existing workflow development and computing environments have limited support for data parallelism patterns. Such limitation makes complex the automation and ability to perform efficient parallel execution on large sets of data, which may significantly slow down the execution of a workflow.

In contrast, peer-to-peer (P2P) systems [9] 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. Popular examples of P2P systems such as Gnutella and BitTorrent 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. To deal with the dynamic behavior of peers that can join and leave the system at any time, they rely on the fact that popular data get massively duplicated.

Initial research on P2P systems has focused on improving the performance of query routing in the unstructured systems which rely on flooding, whereby peers forward messages to their neighbors. This work led to structured solutions based on Distributed Hash Tables (DHT), e.g. CHORD and Pastry, or hybrid solutions with super-peers that index subsets of peers. Another approach is to exploit gossiping protocols, also known as epidemic protocols. Gossiping has been initially proposed to maintain the mutual consistency of replicated data by spreading replica updates to all nodes over the network. It has since been successfully used in P2P networks for data dissemination. Basic gossiping is simple. Each peer has a complete view of the network (i.e. a list of all peers' addresses) and chooses a node at random to spread the request. The main advantage of gossiping is robustness over node failures since, with very high probability, the request is eventually propagated to all nodes in the network. In large P2P networks, however, the basic gossiping model does not scale as maintaining the complete view of the network at each node would generate very heavy communication traffic. A solution to scalable gossiping is by having each peer with only a partial view of the network, e.g. a list of tens of neighbor peers. To gossip a request, a peer chooses at random a peer in its partial view to send it the request. In addition, the peers involved in a gossip exchange their partial views to reflect network changes in their own views. Thus, by continuously refreshing their partial views, nodes can self-organize into randomized overlays which scale up very well.

We claim that a P2P solution is the right solution 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 or to execute very large data-intensive workflow activities (activities that manipulate huge amounts of data), 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 bests of both. In particular, it will enable the clean integration of the users' own computational resources with different clouds.

3.3. Cloud Data Management

Cloud computing encompasses on demand, reliable services provided over the Internet (typically represented as a cloud) with easy access to virtually infinite computing, storage and networking resources. Through very simple Web interfaces and at small incremental cost, users can outsource complex tasks, such as data storage, system administration, or application deployment, to very large data centers operated by cloud providers. Thus, the complexity of managing the software/hardware infrastructure gets shifted from the users' organization to the cloud provider. From a technical point of view, the grand challenge is to support in a cost-effective way the very large scale of the infrastructure which has to manage lots of users and resources with high quality of service.

Cloud customers could move all or part of their information technology (IT) services to the cloud, with the following main benefits:

- **Cost.** The cost for the customer can be greatly reduced since the IT infrastructure does not need to be owned and managed; billing is only based only on resource consumption. For the cloud provider, using a consolidated infrastructure and sharing costs for multiple customers reduces the cost of ownership and operation.
- **Ease of access and use.** The cloud hides the complexity of the IT infrastructure and makes location and distribution transparent. Thus, customers can have access to IT services anytime, and from anywhere with an Internet connection.
- **Quality of Service (QoS).** The operation of the IT infrastructure by a specialized provider that has extensive experience in running very large infrastructures (including its own infrastructure) increases QoS.
- **Elasticity.** The ability to scale resources out, up and down dynamically to accommodate changing conditions is a major advantage. In particular, it makes it easy for customers to deal with sudden increases in loads by simply creating more virtual machines.

However, cloud computing has some drawbacks and not all applications are good candidates for being "cloudified". The major concern is wrt. data security and privacy, and trust in the provider (which may use no so trustful providers to operate). One earlier criticism of cloud computing was that customers get locked in proprietary clouds. It is true that most clouds are proprietary and there are no standards for cloud interoperability. But this is changing with open source cloud software such as Hadoop, an Apache project implementing Google's major cloud services such as Google File System and MapReduce, and Eucalyptus, an open source cloud software infrastructure, which are attracting much interest from research and industry.

There is much more variety in cloud data than in scientific data since there are many different kinds of customers (individuals, SME, large corporations, etc.). However, we can identify common features. Cloud data can be very large, unstructured (e.g. text-based) or semi-structured, and typically append-only (with rare updates). And cloud users and application developers may be in high numbers, but not DBMS experts.

Current cloud data management (NOSQL) solutions typically trade consistency for scalability, simplicity and flexibility. They use a radically different architecture than RDBMS, by exploiting (rather than embedding) a distributed file system such as Google File System (GFS) or Hadoop Distributed File System (HDFS), to store and manage data in a highly fault-tolerant manner. They tend to rely on a more specific data model, e.g. key-value store such Google Bigtable, Hadoop Hbase or Apache CouchDB) with a simple set of operators easy to use from a programming language. For instance, to address the requirements of social network applications, new solutions rely on a graph data model and graph-based operators. User-defined functions also allow for more specific data processing. MapReduce is a good example of generic parallel data processing framework, on top of a distributed file system (GFS or HDFS). It supports a simple data model (sets of (key, value) pairs), which allows user-defined functions (map and reduce). Although quite successful among developers, it is relatively low-level and rigid, leading to custom user code that is hard to maintain and reuse. In Zenith, we exploit or extend these NOSQL technologies to fit our needs for scientific workflow management and scalable data analysis.

3.4. Uncertain Data Management

Data uncertainty is present in many scientific applications. For instance, in the monitoring of plant contamination by INRA teams, sensors generate periodically data which may be uncertain. Instead of ignoring (or correcting) uncertainty, which may generate major errors, we need to manage it rigorously and provide support for querying.

To deal with uncertainty, there are several approaches, e.g. probabilistic, possibilistic, fuzzy logic, etc. The *probabilistic approach* is often used by scientists to model the behavior of their underlying environments. However, in many scientific applications, data management and uncertain query processing are not integrated, i.e. the queries are usually answered using ad-hoc methods after doing manual or semi-automatic statistical treatment on the data which are retrieved from a database. In Zenith, we aim at integrating scientific data management and query processing within one system. This should allow scientists to issue their queries in a query language without thinking about the probabilistic treatment which should be done in background in order to answer the queries. There are two important issues which any PDBMS should address: 1) how to represent a probabilistic database, i.e. data model; 2) how to answer queries using the chosen representation, i.e. query evaluation.

One of the problems on which we focus is *scalable query processing* over uncertain data. A naive solution for evaluating probabilistic queries is to enumerate all possible worlds, i.e. all possible instances of the database, execute the query in each world, and return the possible answers together with their cumulative probabilities. However, this solution can not scale up due to the exponential number of possible worlds which a probabilistic database may have. Thus, the problem is quite challenging, particularly due to exponential number of possibilities that should be considered for evaluating queries. In addition, most of our underlying scientific applications are not centralized; the scientists share part of their data in a *P2P* manner. This distribution of data makes very complicated the processing of probabilistic queries. To develop efficient query processing techniques for distributed scientific applications, we can take advantage of two main distributed technologies: *P2P* and *Cloud*. Our research experience in P2P systems has proved us that we can propose scalable solutions for many data management problems. In addition, we can use the cloud parallel solutions, e.g. MapReduce, to parallelize the task of query processing, when possible, and answer queries of scientists in reasonable execution times. Another challenge for supporting scientific applications is uncertain data integration. In addition to managing the uncertain data for each user, we need to integrate uncertain data from different sources. This requires revisiting traditional data integration in major ways and dealing with the problems of uncertain mediated schema generation and uncertain schema mapping.

3.5. Metadata Integration

Nowdays, scientists can rely on web 2.0 tools to quickly share their data and/or knowledge (e.g. ontologies of the domain knowledge). Therefore, when performing a given study, a scientist would typically need to access and integrate data from many data sources (including public databases). To make high numbers of scientific data sources easily accessible to community members, it is necessary to identifying semantic correspondences between metadata structures or models of the related data sources. The main underlying task is called matching, which is the process of discovering semantic correspondences between metadata structures such as database schema and ontologies. Ontology is a formal and explicit description of a shared conceptualization in term of concepts (i.e., classes, properties and relations). For example, the matching may be used to align gene ontologies or anatomical metadata structures.

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 great autonomy of the underlying data sources, which leads to a large variety of models and formats. The high heterogeneity makes the matching problem very challenging. Furthermore, the number of ontologies and their size grow fastly, so does their diversity and heterogeneity. As a result, schema/ontology matching has become a prominent and challenging topic [4].

3.6. Data Mining

Data mining provides methods to discover new and useful patterns from very large sets of data. These patterns may take different forms, depending on the end-user's request, such as:

- **Frequent itemsets and association rules [1].** In this case, the data is usually a table with a high number of rows and the algorithm extracts correlations between column values. This problem was first motivated by commercial and marketing purposes (*e.g.* discovering frequent correlations between items bought in a shop, which could help selling more). 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 mining, but in this case, the order between events has to be considered. Let us consider the smart-building example again. A frequent sequence, in this case, could say that "in 40% 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". Discovering frequent sequences has become a crucial need in marketing, but also in security (detecting network intrusions for instance) in usage analysis (web usage is one of the main applications) and any domain where data arrive in a specific order (usually given by timestamps).
- **Clustering [12].** The goal of clustering algorithms is to group together data that have similar characteristics, while ensuring that dissimilar data will not be in the same cluster. In our example of smart buildings, we would find clusters of rooms, where offices will be in one category and copy machine rooms in another one because of their characteristics (hours of people presence, number of times lights are turned on and off, etc.).

One of the main problems for data mining methods recently was to deal with data streams. Actually, data mining methods have first been designed for very large data sets where complex algorithms of artificial intelligence were not able to complete within reasonable time responses because of data size. The problem was thus to find a good trade-off between time response and results relevance. The patterns described above well match this trade-off since they both provide interesting knowledge for data analysts and allow algorithm having good time complexity on the number of records. Itemset mining algorithms, for instance, depend more on the number of columns (for a sensor it would be the number of possible items such as temperature, presence, status of lights, etc.) than the number of lines (number of sensors in the network). However, with the ever growing size of data and their production rate, a new kind of data source has recently emerged as data streams. A data stream is a sequence of events arriving at high rate. By "high rate", we usually admit that traditional data mining methods reach their limits and cannot complete in real-time, given the data size. In order to extract knowledge from such streams, a new trade-off had to be found and the data mining community has investigated approximation methods that could allow maintaining a good quality of results for the above patterns extraction.

For scientific data, data mining now has to deal with new and challenging characteristics. First, scientific data is often associated to a level of uncertainty (typically, sensed values have to be associated to the probability that this value is correct or not). Second, scientific data might be extremely large and need cloud computing solutions for their storage and analysis. Eventually, we will have to deal with high dimension and heterogeneous data.

3.7. Content-based Information Retrieval

Today's technologies for searching information in scientific data mainly rely on relational DBMS or text based indexing methods. However, content-based information retrieval has progressed much in the last decade and is now considered as one of the most promising for future search engines. Rather than restricting search to the use of metadata, content-based methods attempt to index, search and browse digital objects by means of signatures describing their actual content. Such methods have been intensively studied in the multimedia community to allow searching the massive amount or raw multimedia documents created every day (e.g. 99% of web data are audio-visual content with very sparse metadata). Successful and scalable content-based

methods have been proposed for searching objects in large image collections or detecting copies in huge video archives. Besides multimedia contents, content-based information retrieval methods recently started to be studied on more diverse data such as medical images, 3D models or even molecular data. Potential applications in scientific data management are numerous. First of all, to allow searching the huge collections of scientific images (earth observation, medical images, botanical images, biology images, etc.) but also to browse large datasets of experimental data (e.g. multisensor data, molecular data or instrumental data). Despite recent progress, scalability remains a major issue, involving complex algorithms (such as similarity search, clustering or supervised retrieval), in high dimensional spaces (up to millions of dimensions) with complex metrics (Lp, Kernels, sets intersections, edit distances, etc.). Most of these algorithms have linear, quadratic or even cubic complexities so that their use at large scale is not affordable without consistent breakthrough. In Zenith, we plan to investigate the following challenges:

- **High-dimensional similarity search**. Whereas many indexing methods were designed in the last 20 years to retrieve efficiently multidimensional data with relatively small dimensions, high-dimensional data have been more challenging due to the well-known dimensionality curse. Only recently have some methods appeared that allow approximate Nearest Neighbors queries in sub-linear time, in particular, Locality Sensitive Hashing methods which offer new theoretical insights in high-dimensional Euclidean spaces and proved the interest of random projections. But there are still some challenging issues that need to be solved including efficient similarity search in any kernel or metric spaces, efficient construction of knn-graphs or relational similarity queries.
- **Large-scale supervised retrieval**. Supervised retrieval aims at retrieving relevant objects in a dataset by providing some positive and/or negative training samples. To solve such task, there has been a focused interest on using Support Vector Machines (SVM) that offer the possibility to construct generalized, non-linear predictors in high-dimensional spaces using small training sets. The prediction time complexity of these methods is usually linear in dataset size. Allowing hyperplane similarity queries in sub-linear time is for example a challenging research issue. A symmetric problem in supervised retrieval consists in retrieving the most relevant object categories that might contain a given query object, providing huge labeled datasets (up to millions of classes and billions of objects) and very few objects per category (from 1 to 100 objects). SVM methods that are formulated as quadratic programming with cubic training time complexity and quadratic space complexity are clearly not usable. Promising solutions to such problems include hybrid supervised-unsupervised methods and supervised hashing methods.
- **P2P content-based retrieval**. Content-based P2P retrieval methods appeared recently as a promising solution to manage masses of data distributed over large social networks, particularly when the data cannot be centralized for privacy or cost reasons (which is often the case in scientific social networks, e.g. botanist social networks). However, current methods are limited to very simple similarity search paradigms. In Zenith, we will consider more advanced P2P content-based retrieval and mining methods such as k-nn graphs construction, large-scale supervised retrieval or multi-source clustering.

COPRIN Project-Team

3. Scientific Foundations

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 [0.5,0.6] 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 than 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 [1], [7], [5]:

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 [11], [19]
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 COPRIN 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).

3.2. Robotics

COPRIN has a long-standing tradition of robotics studies, especially for closed-loop robots [4]. 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 [13], [21] such as workspace and singularity analysis, positioning accuracy [24], trajectory planning, reliability, calibration [33], 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 [28]. The methods that we develop can be used for other robotic problems, see for example the management of uncertainties in aircraft design [34], [10].

Our theoretical work must be validated through experiments that are essential for the sake of credibility. A contrario, experiments will feed theoretical work. Hence COPRIN works with partners on the development of real robots but also develops its own prototypes. We usually develop a new robot prototype every 6 years but since 2008 we have started the development of seven new robot prototypes, mostly related to assistance robotics. Furthermore 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. We intend to make a full use of such devices, especially for assistance purpose

In term of applications we have focused up to now on the development of special machines (machine-tool, ultra-high accuracy positioning device, spatial telescope). Although this activity will be pursued, we have started in 2008 a long-term move toward *service robotics*, i.e. robots that are closer to human activity. In service robotics we are interested in domotics, smart objects, rehabilitation and medical robots [8], [9], [26] and entertainment, that can be regrouped under the name of *assistance robotics* (see section 6.2.1.3). Compared to special machines for which pricing is not an issue (up to a certain point), cost is an important element for assistance robotics. While we plan to develop simple robotic systems using only standard hardware, our work will focus on a different issue: *adaptability*. We aim at providing assistance devices that are adapted to the end-user, its trajectory of life and its environment, are easy to install (because installation uncertainties are taken into account at the design stage), have a low intrusivity and are guaranteed to fulfill a set of requirements.

E-MOTION Project-Team (section vide)

FLOWERS Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

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 grand challenge is thus to be able to build robotic 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" [120]. 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 [90], 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 [97] [122]. The approach of developmental robotics consists in importing and implementing concepts and mechanisms from developmental psychology [102], cognitive linguistics [79], and developmental cognitive neuroscience [93] 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 [76] [110], grounding [88], situatedness [70], self-organization [118] [111], enaction [121], and incremental learning [77].

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 [14]. 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 [102], 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" [72] [84]. 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 [74] [80] [82]. 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 behaviors and curiosity [81] [91] [116]. Based on this, a number of researchers have began in the past few years to build computational implementation of intrinsic motivation [14] [108] [114] [73] [92] [100] [115]. 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 [108][14] [109] [113]. 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 [78] [89] 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 [102]. 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 [71]. 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 [119], [75], motivated by the various mechanisms that allow humans to socially guide a robot [112]. In an interactive context the steps of self-exploration and social guidances are not separated and a robot learns by self exploration and by receiving extra feedback from the social context [119], [94] [101].

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 [104], [106]. We also work on developing perceptual capabilities by gradually segmenting the perceptual space and identifying objects and their characteristics through interaction with the user [48] and robots experiments [43]. 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 [96].

IMARA Project-Team

3. Scientific Foundations

3.1. Vehicle guidance and autonomous navigation

Participants: Fawzi Nashashibi, Evangeline Pollard, Benjamin Lefaudeux, Hao Li, Paulo Lopes Resende, Guillaume Tréhard, Pierre Merdrignac, Zayed Alsayed.

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

Either for driver assistance or for fully automated guided vehicles 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.

This year was the opportunity to focus on two particular topics: SLAMMOT-based techniques and cooperative perception.

3.1.2. 3D environment mapping

Participants: Fawzi Nashashibi, Hao Li, Benjamin Lefaudeux, Paulo Lopes Resende.

In the past few years, we've been focusing on the Disparity map estimation as a mean to obtain dense 3D mapping of the environment. Moreover, many autonomous vehicle navigation systems have adopted stereo vision techniques to construct disparity maps as a basic obstacle detection and avoidance mechanism. Two different approaches where investigated: the Fly algorithm, and the stereo vision for 3D representation.

The Fly Algorithm is an evolutionary optimization applied to stereovision and mobile robotics. Its advantage relies on its precision and its acceptable costs (computation time and resources). In the other approach, originality relies in computing the disparity field by directly formulating the problem as a constrained optimization problem in which a convex objective function is minimized under convex constraints. These constraints arise from prior knowledge and the observed data. The minimization process is carried out over the feasibility set and with a suitable regularization constraint: the Total Variation information, which avoids oscillations while preserving field discontinuities around object edges. Although successfully applied to real-time pedestrian detection using a vehicle mounted stereohead (see LOVe project), this technique could not be used for other robotics applications such as scene modeling, visual SLAM, etc. The need is for a dense 3D representation of the environment obtained with an appropriate precision and acceptable costs (computation time and resources).

Stereo vision is a reliable technique for obtaining a 3D scene representation through a pair of left and right images and it is effective for various tasks in road environments. The most important problem in stereo image processing is to find corresponding pixels from both images, leading to the so-called disparity estimation. Many autonomous vehicle navigation systems have adopted stereo vision techniques to construct disparity maps as a basic obstacle detection and avoidance mechanism. We also worked in the past on an original approach for computing the disparity field by directly formulating the problem as a constrained optimization problem in which a convex objective function is minimized under convex constraints. These constraints arise from prior knowledge and the observed data. The minimization process is carried out over the feasibility set, which corresponds to the intersection of the constraint sets. The construction of convex property sets is based on the various properties of the field to be estimated. In most stereo vision applications, the disparity map should be smooth in homogeneous areas while keeping sharp edges. This can be achieved with the help of a suitable regularization constraint. We propose to use the Total Variation information as a regularization constraint, which avoids oscillations while preserving field discontinuities around object edges.

The algorithm we developed to solve the estimation disparity problem has a block-iterative structure. This allows a wide range of constraints to be easily incorporated, possibly taking advantage of parallel computing architectures. This efficient algorithm allowed us to combine the Total Variation constraint with additional convex constraints so as to smooth homogeneous regions while preserving discontinuities.

Presently, we are currently working on an original stereo-vision based SLAM technique, aimed at reconstructing current surroundings through on-the-fly real time localization of tens of thousands of interest points. This development should also allow detection and tracking of moving objects [1], and is built on linear algebra (through Inria's Eigen library), RANSAC and multi-target tracking techniques, to quote a few.

This technique complements another laser based SLAMMOT technique developed since few years and extensively validated in large scale demonstrations for indoor and outdoor robotics applications. This technique has proved its efficiency in terms of cost, accuracy and reliability.

3.1.3. Cooperative Multi-sensor data fusion

Participants: Fawzi Nashashibi, Hao Li, Evangeline Pollard, Benjamin Lefaudeux, Pierre Merdrignac.

Since data are noisy, inaccurate and can also be unreliable or unsynchronized, the use of data fusion techniques is required in order to provide the most accurate situation assessment as possible to perform the perception task. IMARA team worked a lot on this problem in the past, but is now focusing on collaborative perception approach. Indeed, the use of vehicle-to-vehicle or vehicle-to-infrastructure communications allows an improved on-board reasoning since the decision is made based on an extended perception.

As a direct consequence of the electronics broadly used for vehicular applications, communication technologies are now being adopted as well. In order to limit injuries and to share safety information, research in driving assistance system is now orientating toward the cooperative domain. Advanced Driver Assistance System (ADAS) and Cybercars applications are moving towards vehicle-infrastructure cooperation. In such scenario, information from vehicle based sensors, roadside based sensors and a priori knowledge is generally combined

[1] http://www.youtube.com/watch?v=obH9Z2uOMBI

thanks to wireless communications to build a probabilistic spatio-temporal model of the environment. Depending on the accuracy of such model, very useful applications from driver warning to fully autonomous driving can be performed.

The Collaborative Perception Framework (CPF) is a combined hardware/software approach that permits to see remote information as its own information. Using this approach, a communicant entity can see another remote entity software objects as if it was local, and a sensor object, can see sensor data of others entities as its own sensor data. Last year's developments permitted the development of the basic hardware pieces that ensures the well functioning of the embedded architecture including perception sensors, communication devices and processing tools. The final architecture was relying on the *SensorHub* presented in year 2010 report and demonstrated several times in year 2011 (ITS World Congress, workshop "The automation for urban transport" in La Rochelle...)

Finally, since vehicle localization (ground vehicles) is an important task for intelligent vehicle systems, vehicle cooperation may bring benefits for this task. A new cooperative multi-vehicle localization method using split covariance intersection filter was developed during the year 2012, as well as a cooperative GPS data sharing method.

In the first method, each vehicle estimates its own position using a SLAM approach. In parallel, it estimates a decomposed group state, which is shared with neighboring vehicles; the estimate of the decomposed group state is updated with both the sensor data of the ego-vehicle and the estimates sent from other vehicles; the covariance intersection filter which yields consistent estimates even facing unknown degree of inter-estimate correlation has been used for data fusion.

In the second GPS data sharing method, a new collaborative localization method is proposed. On the assumption that the distance between two communicative vehicles can be calculated with a good precision, cooperative vehicle are considered as additional satellites into the user position calculation by using iterative methods. In order to limit divergence, some filtering process is proposed: Interacting Multiple Model (IMM) is used to guarantee a greater robustness in the user position estimation.

Both methods should be experimentally tested on IMARA veicles in 2013.

3.1.4. Planning and executing vehicle actions

Participants: Plamen Petrov, Joshué Pérez Rastelli, Fawzi Nashashibi, Philippe Morignot, Paulo Lopes Resende, Mohamed Marouf.

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've been focusing on the generation of geometric trajectories as a result of a manoeuvre selection process using grid-based rating technique or fuzzy technique. For high speed vehicles, Partial Motion Planning techniques we tested revealed their limitation 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 DLR's JointSystem demonstrator used in the European project HAVEit as well as in IMARA's electrical vehicle prototype used in the French project ABV. HAVEit was also the opportunity for IMARA 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 IMARA's cybercars. However, for the low speed cybercars that have pre-defined itineraries and basic manoeuvres it was necessary to develop a more adapted planning and control system. Therefore, we've 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, controlling the vehicles include also advanced platooning, automated parking, automated docking, etc. Foreach functionality a dedicated control algorithm was designed (see publication of previous years). Today, IMARA is also investigating the opportunity of fuzzy-based control for specific manoeuvres. First results have been recently obtained for reference trajectory following in roundabouts and normal straight roads.

3.2. V2V and V2I Communications for ITS

Participants: Thierry Ernst, Oyunchimeg Shagdar, Gérard Le Lann, Manabu Tsukada, Thouraya Toukabri, Satoru Noguchi, Ines Ben Jemaa, Mohammad Abu Alhoul, Fawzi Nashashibi, Arnaud de la Fortelle.

Wireless communications is expected to play an important role for road safety, road efficiency, and comfort of road users. Road safety applications often require highly responsive and reliable information exchange between neighboring vehicles in any road density condition. Because the performance of the existing radio communications technology largely degrades with the increase of the node density, the challenge of designing wireless communications for safety applications is enabling reliable communications in highly dense scenarios. Targeting this issue, IMARA has been working on medium access control design and visible light communications especially for highly dense scenarios. The works have been carried out considering vehicles' behavior such as vehicles' merging and platooning.

Unlike many of the road safety applications, the applications regarding road efficiency and comfort of road users, on the other hand, often require connectivity to the Internet. Based on our expertise in both Internet-based communications in the mobility context and in ITS, we are now investigating the use of IPv6 (Internet Protocol version 6 which is going to replace the current version, IPv4, in a few years from now) for vehicular communications, in a combined architecture allowing both V2V and V2I. In the context of IPv6, we have been tackling research issues of combinations of MANET and NEMO and Multihoming in Nested Mobile Networks with Route Optimization.

The wireless channel and topology dynamics are the characteristics that require great research challenge in understanding the dynamics and designing efficient communications mechanisms. Targeting this issue we 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 geocast message delivery, and access point selection.

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

3.2.1. Multihoming in nested mobile networks with route optimization

Participants: Manabu Tsukada, Thierry Ernst.

Network mobility has the particularity of allowing recursive mobility, i.e. where a mobile node is attached to another mobile node (e.g. a PDA is attached to the in-vehicle IP network). This is referred to as nested mobility and brings a number of research issues in terms of routing efficiency. Another issue under such mobility configurations is the availability of multiple paths to the Internet (still in the same example, the PDA has a 3G interface and the in-vehicle network has some dedicated access to the Internet) and its appropriate selection.

3.2.2. Service discovery

Participants: Satoru Noguchi, Thierry Ernst.

Vehicles in a close vicinity need to discover what information can be made available to other vehicles (e.g. road traffic conditions, safety notification for collision avoidance). We are investigating both push and pull approaches and the ability of these mechanisms to scale to a large number of vehicles and services on offer.

3.2.3. Geographic multicast addressing and routing

Participants: Ines Ben Jemaa, Oyunchimeg Shagdar, Thierry Ernst, Arnaud de La Fortelle, Fawzi Nashashibi.

Many ITS applications such as fleet management require multicast data delivery. Existing works on this subject tackle mainly the problems of IP multicasting inside the Internet or geocasting in the VANETs. To enable Internet-based multicast services for VANETs, we introduced a framework that: i) to ensure vehicular multicast group reachability through the infrastructure network, defines a distributed and efficient geographic multicast auto-addressing mechanism and ii) to allow simple and efficient data delivery introduces a simplified approach that locally manages the group membership and distributes the packets among them.

3.2.4. Platooning control using visible light communications

Participants: Mohammad Abu Alhoul, Mohamed Marouf, Oyunchimeg Shagdar, Fawzi Nashashibi.

The main purpose of our research is to propose and test new successful supportive communication technology, which can provide stable and reliable communication between vehicles, especially for the platooning scenario. Although that VLC technology has a short history in comparing with other communication technologies, the infrastructure availability and the presence of the congestion in wireless communication channels are proposing VLC technology as reliable and supportive technology which can takeoff some loads of the wireless radio communication. First objective of this work is develop analytical model of VLC to understand its characteristics and limitation. The second objective of this work is to design vehicle platooning control using VLC. In platooning control, a corporation between control and communication is strongly required in order guarantee the platoon's stability (e.g. string stability problem). For this purpose we work on VLC model platooning scenario, to permit each vehicle the trajectory tracking of the vehicle ahead, altogether with a prescribed inter-vehicle distance and considering all the VLC channel model limitations. The integrated channel model to the main Simulink platooning model will be responsible for deciding the availability of the Line-of-Sight for different trajectory's curvatures, which mean the capability of using light communication between each two vehicles in the platooning queue, at the same time the model will calculate all the required parameters acquired from each vehicle controller.

3.2.5. Access point selection

Participant: Oyunchimeg Shagdar.

While 5.9 GHz radio frequency band is dedicated to ITS applications, there is not much known how the channel and network behave in mobile scenarios. In this work we theoretically and experimentally study the radio channel characteristics in vehicular networks, especially the radio quality and bandwidth availability. Based on our study we develop access point selection method to achieve high speed V2I communications.

3.3. Automated driving, intelligent vehicular networks, and safety

Participant: Gérard Le Lann.

Intelligent vehicular networks (IVNs) are one constituent of ITS. IVNs encompass "clusters", platoons and vehicular ad-hoc networks comprising automated and cooperative vehicles. A basic principle that underlies our work is minimal reliance on road-side infrastructures for solving those open problems arising with IVNs. For example, V2V communications only are considered. Trivially, if one can solve a problem P considering V2V communications only, then P is solved with the help of V2I communications, whereas the converse is not true. Moreover, safety in the course of risk-prone maneuvers is our central concern. Since safety-critical scenarios may develop anytime anywhere, it is impossible to assume that there is always a road-side unit in the vicinity of those vehicles involved in a hazardous situation.

3.3.1. Cohorts and groups – Novel constructs for safe IVNs

The automated driving function rests on two radically different sets of solutions, one set encompassing signal processing and robotics (SPR), the other one encompassing vehicular communications and networking (VCN). In addition to being used for backing a failing SPR solution, VCN solutions have been originally proposed for "augmenting" the capabilities offered by SPR solutions, which are line-of-sight technologies, i.e. limited by obstacles. Since V2V omnidirectional radio communications that are being standardized (IEEE 802.11p / WAVE) have ranges in the order of 250 m, it is interesting to prefix risk-prone maneuvers with the exchange of SC messages. Roles being assigned prior to initiating physical maneuvers, the SPR solutions are invoked under favorable conditions, safer than when vehicles have not agreed on "what to do" ahead of time.

VCN solutions shall belong to two categories: V2V omnidirectional (360°) communications and unidirectional communications, implemented out of very-short range antennas of very small beamwidth. This has led to the concept of neighbor-to-neighbor (N2N) communications, whereby vehicles following each other on a given lane can exchange periodic beacons and event-driven messages.

Vehicle motions on roads and highways obey two different regimes. First, stationary regimes, where inter-vehicular spacing, acceleration and deceleration rates (among other parameters), match specified bounds. This, combined with N2N communications, has led to the concept of cohorts, where safety is not at stake provided that no violation of bounds occurs. Second, transitory regimes, where some of these bounds are violated (e.g., sudden braking – the "brick wall" paradigm), or where vehicles undertake risk-prone maneuvers such as lane changes, resulting into SC scenarios. Reasoning about SC scenarios has led to the concept of groups. Cohorts and groups have been introduced in [7] and [31].

3.3.2. Cohorts, N2N communications, and safety in the presence of telemetry failures

In [7] and [31], we show how periodic N2N beaconing serves to withstand failures of directional telemetry devices. Worst-case bounds on safe inter-vehicular spacing are established analytically (simulations cannot be used for establishing worst-case bounds). A result of practical interest is the ability to answer the following question: "vehicles move at high speed in a cohort formation; if in a platoon formation, spacing would be in the order of 3 m; what is the additional safe spacing in a cohort?" With a N2N beaconing period in the range of 100-200 ms, the additional spacing is much less than 1 m. Failure of a N2N communication link translates into a cohort split, one of the vehicles impaired becoming the tail of a cohort, and its (impaired) follower becoming the head of a newly formed cohort. The number of vehicles in a cohort has an upper bound, and the inter-cohort spacing has a lower bound.

3.3.3. Groups, cohorts, and fast reliable V2V Xcasting in the presence of message losses

Demonstrating safety involves establishing strict timeliness ("real time") properties under worst-case conditions (traffic density, failure rates, radio interference ranges). As regards V2V message passing, this requirement translates into two major problems:

- TBD: time-bounded delivery of V2V messages exchanged among vehicles that undertake SC maneuvers, despite high message loss ratios.
- TBA: time-bounded access to a radio channel in open ad hoc, highly mobile, networks of vehicles, some vehicles undertaking SC maneuvers, despite high contention.

Groups and cohorts have proved to be essential constructs for devising a solution for problem TBD. Vehicles involved in a SC scenario form a group where a 3-way handshake is unfolded so as to reach an agreement regarding roles and adjusted motions. A 3-way handshake consists in 3 rounds of V2V Xcasting of SC messages, round 1 being a Geocast, round 2 being a Convergecast, and round 3 being a Multicast. Worst-case time bound for completing a 3-way handshake successfully is in the order of 200 ms, under worst-case conditions. It is well known that message losses are the dominant cause of failures in mobile wireless networks, which raises the following problem with the Xcasting of SC messages. If acknowledgments are not used, it is impossible to predict probabilities for successful deliveries, which is antagonistic with demonstrating safety. Asking for acknowledgments is a non solution. Firstly, by definition, vehicles that are to be reached by a Geocast are unknown to a sender. How can a sender know which acknowledgments to wait for? Secondly, repeating a SC message that has been lost on a radio channel does not necessarily increase chances of successful delivery. Indeed, radio interferences (causing the first transmission loss) may well last longer than 200 ms (or seconds). To be realistic, one is led to consider a novel and extremely powerful (adversary) failure model (denoted Ω), namely the restricted unbounded omission model, whereby messages meant to circulate on f out of n radio links are "erased" by the adversary (the same f links), ad infinitum. Moreover, we have assumed message loss ratios f/n as high as 2/3. This is the setting we have considered in [49], where we present a solution for the fast (less than 200 ms) reliable (in the presence of Ω) multipoint communications problem TBD. The solution consists in a suite of Xcast protocols (the Zebra suite) and proxy sets built out of cohorts. Analytical expressions are given for the worst-case time bounds for each of the Zebra protocols.

Surprisingly, while not being originally devised to that end, it turns out that cohorts and groups are essential cornerstones for solving open problem TBA.

3.4. Managing the system (via probabilistic modeling)

Participants: Guy Fayolle, Cyril Furtlehner, Yufei Han, Arnaud de La Fortelle, Jean-Marc Lasgouttes, Victorin Martin.

The research on the management of the transportation system is a natural continuation of the research of the Preval team, which joined IMARA in 2007. For many years, the members of this team (and of its ancestor Meval) have been working on understanding random systems of various origins, mainly through the definition and solution of mathematical models. The traffic modeling field is very fertile in difficult problems, and it has been part of the activities of the members of Preval since the times of the Praxitèle project.

Following this tradition, the roadmap of the group is to pursue basic research on probabilistic modeling with a clear slant on applications related to LaRA activities. A particular effort is made to publicize our results among the traffic analysis community, and to implement our algorithms whenever it makes sense to use them in traffic management. Of course, as aforementioned, these activities in no way preclude the continuation of the methodological work achieved in the group for many years in various fields: random walks in Z_+^n ([1], [2], [5]), large deviations, birth and death processes on trees, particle systems. The reader is therefore encouraged to read the recent activity reports for the Preval team for more details.

In practice, the group explores the links between large random systems and statistical physics, since this approach proves very powerful, both for macroscopic (fleet management [4]) and microscopic (car-level description of traffic, formation of jams) analysis. The general setting is mathematical modeling of large systems (mostly stochastic), without any a priori restriction: networks [3], random graphs or even objects coming from biology. When the size or the volume of those structures grows (this corresponds to the so-called thermodynamical limit), one aims at establishing a classification based on criteria of a twofold nature: quantitative (performance, throughput, etc) and qualitative (stability, asymptotic behavior, phase transition, complexity).

3.4.1. Exclusion processes

One of the simplest basic (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.

Most of these generalizations lead to models that are obviously difficult to solve and require upstream theoretical studies. Some of them models have already been investigated by members of the group, and they are part of wide ongoing research.

3.4.2. Message passing algorithms

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. Using the Ising model, together with the Belief Propagation algorithm very popular in the computer science community, 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 properties of the Bethe approximation of Ising models:

- determine the effect of the various variants of BP (in terms of normalization or changes to the Bethe free energy) on the fixed points and their stability;

- find the best way to inject real-valued data in an Ising model with binary variables;
- build macroscopic variables that measure the overall state of the underlying graph, in order to improve the local propagation of information;
- make the underlying model as sparse as possible, in order to improve BP convergence and quality.

LAGADIC Project-Team

3. Scientific Foundations

3.1. Visual servoing

Basically, visual servoing techniques consist in using the data provided by one or several cameras in order to control the motions of a dynamic system [1]. Such systems are usually robot arms, or mobile robots, but can also be virtual robots, or even a virtual camera. A large variety of positioning tasks, or mobile target tracking, can be implemented by controlling from one to all the degrees of freedom of the system. Whatever the sensor configuration, which can vary from one on-board camera on the robot end-effector to several free-standing cameras, a set of visual features has to be selected at best from the image measurements available, allowing to control the desired degrees of freedom. A control law has also to be designed so that these visual features $\mathbf{s}(t)$ reach a desired value $\mathbf{s}^*$, defining a correct realization of the task. A desired planned trajectory $\mathbf{s}^*(t)$ can also be tracked. The control principle is thus to regulate to zero the error vector $\mathbf{s}(t) - \mathbf{s}^*(t)$. With a vision sensor providing 2D measurements, potential visual features are numerous, since 2D data (coordinates of feature points in the image, moments, ...) as well as 3D data provided by a localization algorithm exploiting the extracted 2D features can be considered. It is also possible to combine 2D and 3D visual features to take the advantages of each approach while avoiding their respective drawbacks.

More precisely, a set $\mathbf{s}$ of k visual features can be taken into account in a visual servoing scheme if it can be written:

$$\mathbf{s} = \mathbf{s}(\mathbf{x}(\mathbf{p}(t)), \mathbf{a}) \tag{2}$$

where $\mathbf{p}(t)$ describes the pose at the instant t between the camera frame and the target frame, $\mathbf{x}$ the image measurements, and $\mathbf{a}$ a set of parameters encoding a potential additional knowledge, if available (such as for instance a coarse approximation of the camera calibration parameters, or the 3D model of the target in some cases).

The time variation of $\mathbf{s}$ can be linked to the relative instantaneous velocity $\mathbf{v}$ between the camera and the scene:

$$\dot{\mathbf{s}} = \frac{\partial \mathbf{s}}{\partial \mathbf{p}} \dot{\mathbf{p}} = \mathbf{L}_{\mathbf{s}} \mathbf{v} \tag{3}$$

where $\mathbf{L}_{\mathbf{s}}$ is the interaction matrix related to $\mathbf{s}$. This interaction matrix plays an essential role. Indeed, if we consider for instance an eye-in-hand system and the camera velocity as input of the robot controller, we obtain when the control law is designed to try to obtain an exponential decoupled decrease of the error:

$$\mathbf{v}_c = -\lambda \widehat{\mathbf{L}_{\mathbf{s}}}^{+} (\mathbf{s} - \mathbf{s}^*) - \widehat{\mathbf{L}_{\mathbf{s}}}^{+} \widehat{\frac{\partial \mathbf{s}}{\partial t}} \tag{4}$$

where λ is a proportional gain that has to be tuned to minimize the time-to-convergence, $\widehat{\mathbf{L}_{\mathbf{s}}}^{+}$ is the pseudo-inverse of a model or an approximation of the interaction matrix, and $\widehat{\frac{\partial \mathbf{s}}{\partial t}}$ an estimation of the features velocity due to a possible own object motion.

From the selected visual features and the corresponding interaction matrix, the behavior of the system will have particular properties as for stability, robustness with respect to noise or to calibration errors, robot 3D trajectory, etc. Usually, the interaction matrix is composed of highly non linear terms and does not present any decoupling properties. This is generally the case when $\mathbf{s}$ is directly chosen as $\mathbf{x}$. In some cases, it may lead to inadequate robot trajectories or even motions impossible to realize, local minimum, tasks singularities, etc. It is thus extremely important to design adequate visual features for each robot task or application, the ideal case (very difficult to obtain) being when the corresponding interaction matrix is constant, leading to a simple linear control system. To conclude in few words, **visual servoing is basically a non linear control problem. Our Holy Grail quest is to transform it into a linear control problem.**

Furthermore, embedding visual servoing in the task function approach allows solving efficiently the redundancy problems that appear when the visual task does not constrain all the degrees of freedom of the system. It is then possible to realize simultaneously the visual task and secondary tasks such as visual inspection, or joint limits or singularities avoidance. This formalism can also be used for tasks sequencing purposes in order to deal with high level complex applications.

3.2. Visual tracking

Elaboration of object tracking algorithms in image sequences is an important issue for researches and applications related to visual servoing and more generally for robot vision. A robust extraction and real time spatio-temporal tracking process of visual cues is indeed one of the keys to success of a visual servoing task. If fiducial markers may still be useful to validate theoretical aspects in modeling and control, natural scenes with non cooperative objects and subject to various illumination conditions have to be considered for addressing large scale realistic applications.

Most of the available tracking methods can be divided into two main classes: feature-based and model-based. The former approach focuses on tracking 2D features such as geometrical primitives (points, segments, circles,...), object contours, regions of interest...The latter explicitly uses a model of the tracked objects. This can be either a 3D model or a 2D template of the object. This second class of methods usually provides a more robust solution. Indeed, the main advantage of the model-based methods is that the knowledge about the scene allows improving tracking robustness and performance, by being able to predict hidden movements of the object, detect partial occlusions and acts to reduce the effects of outliers. The challenge is to build algorithms that are fast and robust enough to meet our applications requirements. Therefore, even if we still consider 2D features tracking in some cases, our researches mainly focus on real-time 3D model-based tracking, since these approaches are very accurate, robust, and well adapted to any class of visual servoing schemes. Furthermore, they also meet the requirements of other classes of application, such as augmented reality.

3.3. Slam

Most of the applications involving mobile robotic systems (ground vehicles, aerial robots, automated submarines,...) require a reliable localization of the robot in its environment. A challenging problem is when neither the robot localization nor the map is known. Localization and mapping must then be considered concurrently. This problem is known as Simultaneous Localization And Mapping (Slam). In this case, the robot moves from an unknown location in an unknown environment and proceeds to incrementally build up a navigation map of the environment, while simultaneously using this map to update its estimated position.

Nevertheless, solving the Slam problem is not sufficient for guaranteeing an autonomous and safe navigation. The choice of the representation of the map is, of course, essential. The representation has to support the different levels of the navigation process: motion planning, motion execution and collision avoidance and, at the global level, the definition of an optimal strategy of displacement. The original formulation of the Slam problem is purely metric (since it basically consists in estimating the Cartesian situations of the robot and a set of landmarks), and it does not involve complex representations of the environment. However, it is now well recognized that **several complementary representations are needed to perform exploration, navigation, mapping, and control tasks successfully. We propose to use composite models of the environment that**

mix topological, metric, and grid-based representations. Each type of representation is well adapted to a particular aspect of autonomous navigation: the metric model allows one to locate the robot precisely and plan Cartesian paths, the topological model captures the accessibility of different sites in the environment and allows a coarse localization, and finally the grid representation is useful to characterize the free space and design potential functions used for reactive obstacle avoidance. However, ensuring the consistency of these various representations during the robot exploration, and merging observations acquired from different viewpoints by several cooperative robots, are difficult problems. This is particularly true when different sensing modalities are involved. New studies to derive efficient algorithms for manipulating the hybrid representations (merging, updating, filtering...) while preserving their consistency are needed.

AYIN Team

3. Scientific Foundations

3.1. Probabilistic approaches

Following a Bayesian methodology as far as possible, probabilistic models are used within the AYIN team for three purposes: to describe the class of images to be expected from any given scene, to describe prior knowledge about the scene and to incorporate specific constraints. The models used in AYIN fall into the following two classes.

3.1.1. Markov random fields

Markov random fields were introduced to image processing in the Eighties, and were quickly applied to the full range of inverse problems in computer vision. They owe their popularity to their flexible and intuitive nature, which makes them an ideal modelling tool, and to the existence of standard and easy-to-implement algorithms for their solution. In the AYIN team, attention is focused on their use in image modelling, in particular of textures; on the development of improved prior models for segmentation; and on the lightening of the heavy computational load traditionally associated with these techniques, in particular via the study of varieties of hierarchical random fields.

3.1.2. Stochastic geometry

One of the grand challenges of computer vision and image processing is the expression and use of prior geometric information. For satellite and aerial imagery, this problem has become increasingly important as the increasing resolution of the data results in the necessity to model geometric structures hitherto invisible. One of the most promising approaches to the inclusion of this type of information is stochastic geometry, which is an important line of research in the AYIN team. Instead of defining probabilities for different types of image, probabilities are defined for configurations of an indeterminate number of interacting, parameterized objects located in the image. Such probability distribution are called ‘marked point processes’. Such processes have been recently applied to skin care problems.

3.2. Parameter estimation

One of the most important problems studied in the AYIN team is how to estimate the parameters that appear in the models. For probabilistic models, the problem is quite easily framed, but is not necessarily easy to solve, particularly in the case when it is necessary to extract simultaneously both the information of interest and the parameters from the data.

3.3. Hierarchical models

Another line of research in the AYIN team concerns development of graph-based, in particular, hierarchical models for very high resolution image analysis and classification. A specific hierarchical model recently developed in AYIN represents an image by a forest structure, where leaf nodes represent image regions at the finest level of partition, while other nodes correspond to image regions at the coarser levels of partitions. The AYIN team is interested in developing multi-feature models of image regions as an ensemble of spectral, texture, geometrical and classification features, and establishing new criteria for comparing image regions. Recent research concerns extension of hierarchical models to a temporal dimension, for analyzing multitemporal data series.

IMEDIA2 Team

3. Scientific Foundations

3.1. Introduction

We group the existing problems in the domain of content-based image indexing and retrieval in the following themes: image indexing and efficient search in image collections, pattern recognition and personalization. In the following we give a short introduction to each of these themes.

3.2. Modeling, construction and structuring of the feature space

Participants: Vera Bakic, Nozha Boujemaa, Esma Elghoul, Hervé Goëau, Amel Hamzaoui, Sofiene Mouine, Olfa Mzoughi, Saloua Ouertani-Litayem, Mohamed Riadh Trad, Anne Verroust-Blondet, Itheri Yahiaoui, Zahraa Yasseen.

The goal of IMEDIA2 team is to provide the user with the ability to do content-based search into image databases in a way that is both intelligent and intuitive to the users. When formulated in concrete terms, this problem gives birth to several mathematical and algorithmic challenges.

To represent the content of an image, we are looking for a representation that is both compact (less data and more semantics), relevant (with respect to the visual content and the users) and fast to compute and compare. The choice of the feature space consists in selecting the significant *features*, the *descriptors* for those features and eventually the encoding of those descriptors as image *signatures*.

We deal both with generic databases, in which images are heterogeneous (for instance, search of Internet images), and with specific databases, dedicated to a specific application field. The specific databases are usually provided with a ground-truth and have an homogeneous content (leaf images, for example)

We must not only distinguish generic and specific signatures, but also local and global ones. They correspond respectively to queries concerning parts of pictures or entire pictures. In this case, we can again distinguish approximate and precise queries. In the latter case one has to be provided with various descriptions of parts of images, as well as with means to specify them as regions of interest. In particular, we have to define both global and local similarity measures.

When the computation of signatures is over, the image database is finally encoded as a set of points in a high-dimensional space: the feature space.
A second step in the construction of the index can be valuable when dealing with very high-dimensional feature spaces. It consists in pre-structuring the set of signatures and storing it efficiently, in order to reduce access time for future queries (trade-off between the access time and the cost of storage). In this second step, we have to address problems that have been dealt with for some time in the database community, but arise here in a new context: image databases. Today's scalability issues already put brake on growth of multi-media search engines. The space created by the massive amounts of existing multimedia files greatly exceeds the area searched by today's major engines. Consistent breakthroughs are therefore urgent if we don't want to be lost in data space in ten years. We believe that reducing algorithm complexity remains the main key. Whatever the efficiency of the implementation or the use of powerful hardware and distributed architectures, the ability of an algorithm to scale-up is strongly related to its time and space complexities. Nowadays, efficient multimedia search engines rely on various high level tasks such as content-based search, navigation, knowledge discovery, personalization, collaborative filtering or social tagging. They involve complex algorithms such as similarity search, clustering or machine learning, on heterogeneous data, and with heterogeneous metrics. Some of them still have quadratic and even cubic complexities so that their use in the large scale is not affordable if no fundamental research is performed to reduce their complexities. In this way, efficient and generic high-dimensional similarity search structures are essential for building scalable content-based search systems. Efficient search requires a specific structuring of the feature space (multidimensional indexing, where indexing is understood as data structure) for accelerating the access to collections that are too large for the central memory.

3.3. Pattern recognition and statistical learning

Participants: Nozha Boujemaa, Michel Crucianu, Donald Geman, Wajih Ouertani, Asma Rejeb Sfar.

Statistical learning and classification methods are of central interest for content-based image retrieval. We consider here both supervised and unsupervised methods. Depending on our knowledge of the contents of a database, we may or may not be provided with a set of *labeled training examples*. For the detection of *known* objects, methods based on hierarchies of classifiers have been investigated. In this context, face detection was a main topic, as it can automatically provide a high-level semantic information about video streams. For a collection of pictures whose content is unknown, e.g. in a navigation scenario, we are investigating techniques that adaptively identify homogeneous clusters of images, which represent a challenging problem due to feature space configuration.

Object detection is the most straightforward solution to the challenge of content-based image indexing. Classical approaches (artificial neural networks, support vector machines, etc.) are based on induction, they construct generalization rules from training examples. The generalization error of these techniques can be controlled, given the complexity of the models considered and the size of the training set.

Our research on object detection addresses the design of invariant kernels and algorithmically efficient solutions as well as boosting method for similarity learning. We have developed several algorithms for face detection based on a hierarchical combination of simple two-class classifiers. Such architectures concentrate the computation on ambiguous parts of the scene and achieve error rates as good as those of far more expensive techniques.

Unsupervised clustering techniques automatically define categories and are for us a matter of visual knowledge discovery. We need them in order to:

- Solve the "page zero" problem by generating a visual summary of a database that takes into account all the available signatures together.
- Perform image segmentation by clustering local image descriptors.
- Structure and sort out the signature space for either global or local signatures, allowing a hierarchical search that is necessarily more efficient as it only requires to "scan" the representatives of the resulting clusters.

Given the complexity of the feature spaces we are considering, this is a very difficult task. Noise and class overlap challenge the estimation of the parameters for each cluster. The main aspects that define the clustering process and inevitably influence the quality of the result are the clustering criterion, the similarity measure and the data model.

LEAR Project-Team

3. Scientific Foundations

3.1. Image features and descriptors and robust correspondence

Reliable image features are a crucial component of any visual recognition system. Despite much progress, research is still needed in this area. Elementary features and descriptors suffice for a few applications, but their lack of robustness and invariance puts a heavy burden on the learning method and the training data, ultimately limiting the performance that can be achieved. More sophisticated descriptors allow better inter-class separation and hence simpler learning methods, potentially enabling generalization from just a few examples and avoiding the need for large, carefully engineered training databases.

The feature and descriptor families that we advocate typically share several basic properties:

- **Locality and redundancy:** For resistance to variable intra-class geometry, occlusions, changes of viewpoint and background, and individual feature extraction failures, descriptors should have relatively small spatial support and there should be many of them in each image. Schemes based on collections of image patches or fragments are more robust and better adapted to object-level queries than global whole-image descriptors. A typical scheme thus selects an appropriate set of image fragments, calculates robust appearance descriptors over each of these, and uses the resulting collection of descriptors as a characterization of the image or object (a "bag-of-features" approach – see below).
- **Photometric and geometric invariance:** Features and descriptors must be sufficiently invariant to changes of illumination and image quantization and to variations of local image geometry induced by changes of viewpoint, viewing distance, image sampling and by local intra-class variability. In practice, for local features geometric invariance is usually approximated by invariance to Euclidean, similarity or affine transforms of the local image.
- **Repeatability and salience:** Fragments are not very useful unless they can be extracted reliably and found again in other images. Rather than using dense sets of fragments, we often focus on local descriptors based at particularly salient points – "keypoints" or "points of interest". This gives a sparser and thus potentially more efficient representation, and one that can be constructed automatically in a preprocessing step. To be useful, such points must be accurately relocalizable in other images, with respect to both position and scale.
- **Informativeness:** Notwithstanding the above forms of robustness, descriptors must also be informative in the sense that they are rich sources of information about image content that can easily be exploited in scene characterization and object recognition tasks. Images contain a lot of variety so high dimensional descriptions are required. The useful information should also be manifest, not hidden in fine details or obscure high-order correlations. In particular, image formation is essentially a spatial process, so relative position information needs to be made explicit, e.g. using local feature or context style descriptors.

Partly owing to our own investigations, features and descriptors with some or all of these properties have become popular choices for visual correspondence and recognition, particularly when large changes of viewpoint may occur. One notable success to which we contributed is the rise of "bag-of-features" methods for visual object recognition. These characterize images by their (suitably quantized or parametrized) global distributions of local descriptors in descriptor space. The representation evolved from texton based methods in texture analysis. Despite the fact that it does not (explicitly) encode much spatial structure, it turns out to be surprisingly powerful for recognizing more structural object categories.

Our current research on local features is focused on creating detectors and descriptors that are better adapted to describe object classes, on incorporating spatial neighborhood and region constraints to improve informativeness relative to the bag-of-features approach, and on extending the scheme to cover different kinds of locality. Current research also includes the development and evaluation of local descriptors for video, and associated detectors for spatio-temporal content.

3.2. Statistical modeling and machine learning for image analysis

We are interested in learning and statistics mainly as technologies for attacking difficult vision problems, so we take an eclectic approach, using a broad spectrum of techniques ranging from classical statistical generative and discriminative models to modern kernel, margin and boosting based approaches. Hereafter we enumerate a set of approaches that address some problems encountered in this context.

- Parameter-rich models and limited training data are the norm in vision, so overfitting needs to be estimated by cross-validation, information criteria or capacity bounds and controlled by regularization, model and feature selection.
- Visual descriptors tend to be high dimensional and redundant, so we often preprocess data to reduce it to more manageable terms using dimensionality reduction techniques including PCA and its non-linear variants, latent structure methods such as Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA), and manifold methods such as Isomap/LLE.
- To capture the shapes of complex probability distributions over high dimensional descriptor spaces, we either fit mixture models and similar structured semi-parametric probability models, or reduce them to histograms using vector quantization techniques such as K-means or latent semantic structure models.
- Missing data is common owing to unknown class labels, feature detection failures, occlusions and intra-class variability, so we need to use data completion techniques based on variational methods, belief propagation or MCMC sampling.
- Weakly labeled data is also common – for example one may be told that a training image contains an object of some class, but not where the object is in the image – and variants of unsupervised, semi-supervised and co-training are useful for handling this. In general, it is expensive and tedious to label large numbers of training images so less supervised data mining style methods are an area that needs to be developed.
- On the discriminative side, machine learning techniques such as Support Vector Machines, Relevance Vector Machines, and Boosting, are used to produce flexible classifiers and regression methods based on visual descriptors.
- Visual categories have a rich nested structure, so techniques that handle large numbers of classes and nested classes are especially interesting to us.
- Images and videos contain huge amounts of data, so we need to use algorithms suited to large-scale learning problems.

3.3. Visual recognition and content analysis

Current progress in visual recognition shows that combining advanced image descriptors with modern learning and statistical modeling techniques is producing significant advances. We believe that, taken together and tightly integrated, these techniques have the potential to make visual recognition a mainstream technology that is regularly used in applications ranging from visual navigation through image and video databases to human-computer interfaces and smart rooms.

The recognition strategies that we advocate make full use of the robustness of our invariant image features and the richness of the corresponding descriptors to provide a vocabulary of base features that already goes a long way towards characterizing the category being recognized. Trying to learn everything from scratch using simpler, non-invariant features would require far too much data: good learning cannot easily make up for bad features. The final classifier is thus responsible “only” for extending the base results to larger amounts of intra-class and viewpoint variation and for capturing higher-order correlations that are needed to fine tune the performance.

That said, learning is not restricted to the classifier and feature sets can not be designed in isolation. We advocate an end-to-end engineering approach in which each stage of the processing chain combines learning with well-informed design and exploitation of statistical and structural domain models. Each stage is thoroughly tested to quantify and optimize its performance, thus generating or selecting robust and informative features, descriptors and comparison metrics, squeezing out redundancy and bringing out informativeness.

MAGRIT Project-Team

3. Scientific Foundations

3.1. Camera calibration and registration

One of the most basic problems currently limiting Augmented Reality 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, and the user should walk anywhere he pleases.

For several years, the Magrit project has been aiming at developing on-line and marker-less methods for camera pose computation. We have especially proposed a real-time system for camera tracking designed for indoor scenes [1]. The main difficulty with on-line tracking is to ensure robustness of the process. For off-line processes, robustness is achieved by using spatial and temporal coherence of the considered sequence through move-matching techniques. To get robustness for open-loop systems, we have developed a method which combines the advantage of move-matching methods and model-based methods by using a piecewise-planar model of the environment. This methodology can be used in a wide variety of environments: indoor scenes, urban scenes .... We are also concerned with the development of methods for camera stabilization. Indeed, statistical fluctuations in the viewpoint computations lead to unpleasant jittering or sliding effects, especially when the camera motion is small. We have proved that the use of model selection allows us to noticeably improve the visual impression and to reduce drift over time.

The success of pose computation largely depends on the quality of the matching stage over the sequence. Research are conducted in the team on the use of probabilistic methods to establish robust correspondences of features over time. The use of *a contrario* decision is under investigation to achieve this aim [3]. We especially address the complex case of matching in scenes with repeated patterns which are common in urban scenes. We also consider learning based techniques to improve the robustness of the matching stage.

Another way 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. Each technology approach has limitations: on the one hand, rapid head motions cause image features to undergo large motion between frames that may cause visual tracking to fail. On the other hand, inertial sensors response is largely independent from the user's motion but their accuracy is bad and their response is sensitive to metallic objects in the scene. In past works [1], we have proposed a system that makes an inertial sensor cooperate with the camera-based system in order to improve the robustness of the AR system to abrupt motions of the users, especially head motions. This work contributes to the reduction of the constraints on the users and the need to carefully control the environment during an AR application. Ongoing research on such hybrid systems are under consideration in our team with the aim to improve the accuracy of reconstruction techniques as well as to obtain dynamic models of organs in medical applications.

Finally, it must be noted that the registration problem must be addressed from the specific point of view of augmented reality: the success and the acceptance of an AR application does not only depend on the accuracy of the pose computation but also on the visual impression of the augmented scene. The search for the best compromise between accuracy and perception is therefore an important issue in this project. This research topic has been addressed in our project both in classical AR and in medical imaging in order to choose the camera model, including intrinsic parameters, which describes at best the considered camera.

3.2. Scene 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 occlusion and to compute light reflexions between the real and the virtual objects. Unlike pose computation which has to be computed in a sequential way, scene modeling can be considered as an off-line or an on-line problem according to the application. Within the team we have developed interactive in-situ modeling techniques dedicated to classical AR applications. We also developed off-line multimodal techniques dedicated to AR medical applications.

In-situ modeling
Most automatic techniques aim at reconstructing a sparse and thus unstructured set of points of the scene. Such models are obviously not appropriate to perform interaction with the scene. In addition, they are incomplete in the sense that they may omit features which are important for the accuracy of the pose recovered from 2D/3D correspondences. We have thus investigated interactive techniques with the aim of obtaining reliable and structured models of the scene. The goal of our approach is to develop immersive and intuitive interaction techniques which allow for scene modeling during the application [7].

Multimodal modeling 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 are acquired on the patient or possibly on the operating room through sensing technologies or various image acquisitions. 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's view of the environment.

Methods for multimodal modeling are strongly dependent on the image modalities and the organ specificities. We thus only address a restricted number of medical applications –interventional neuro-radiology and the Augmented Head project– 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 surgeon's training or patient's re-education/learning.

One of our main applications is about neuroradiology. For the last 15 years, we have been working in close collaboration with the neuroradiology laboratory (CHU-University Hospital of Nancy) and GE Healthcare. As several imaging modalities are now available in a intraoperative context (2D and 3D angiography, MRI, ...), our aim is to develop a multi-modality framework to help 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. An accurate definition of the target is a parameter of great importance for the success of the treatment. 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 help 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 Shacra EPI, we have proposed new methods for implicit modeling of the aneurysms with the aim of obtaining near real time simulation of the coil deployment in the aneurysm [4]. Multi-modality techniques for reconstruction are also considered within the european ASPI project, the aim of which is to build a dynamic model of the vocal tract from various images modalities (MRI, ultrasound, video) and magnetic sensors.

MORPHEO Team

3. Scientific Foundations

3.1. Shape Acquisition

Recovering shapes from images is a fundamental task in computer vision. Applications are numerous and include, in particular, 3D modeling applications and mixed reality applications where real shapes are mixed with virtual environments. The problem faced here is to recover shape information such as surfaces from image information. A tremendous research effort has been made in the past to solve this problem in the static case and a number of solutions had been proposed. However, a fundamental issue still to be addressed is the recovery of full shape models with possibly evolving topologies using time sequence information. The main difficulties are precision, robustness of computed shapes as well as consistency of these models over time. Additional difficulties include the integration of multi-modality sensors as well as real-time applications.

3.2. Bayesian Inference

Acquisition of 4D Models can often be conveniently formulated as a Bayesian estimation or learning problem. Various generative and graphical models can be proposed for the problems of occupancy estimation, 3D surface tracking in a time sequence, and motion segmentation. The idea of these generative models is to predict the noisy measurements (e.g. pixel values, measured 3D points or speed quantities) from a set of parameters describing the unobserved scene state, which in turn can be estimated using Bayes' rule to solve the inverse problem. The advantages of this type of modeling are numerous, as they enable to model the noisy relationships between observed and unknown quantities specific to the problem, deal with outliers, and allow to efficiently account for various types of priors about the scene and its semantics. Sensor models for different modalities can also easily be seamlessly integrated and jointly used, which remains central to our goals.

Since the acquisition problems often involve a large number of variables, a key challenge is to exhibit models which correctly account for the observed phenoma, while keeping reasonable estimation times, sometimes with a real-time objective. Maximum likelihood / maximum a posteriori estimation and approximate inference techniques, such as Expectation Maximization, Variational Bayesian inference, or Belief Propagation, are useful tools to keep the estimation tractable. While 3D acquisition has been extensively explored, the research community faces many open challenges in how to model and specify more efficient priors for 4D acquisition and temporal evolution.

3.3. Spectral Geometry

Spectral geometry processing consists of designing methods to process and transform geometric objects that operate in frequency space. This is similar to what is done in signal processing and image processing where signals are transposed into an alternative frequency space. The main interest is that a 3D shape is mapped into a spectral space in a pose-independent way. In other words, if the deformations undergone by the shape are metric preserving, all the meshes are mapped to a similar place in spectral space. Recovering the coherence between shapes is then simplified, and the spectral space acts as a "common language" for all shapes that facilitates the computation of a one-to-one mapping between pairs of meshes and hence their comparisons. However, several difficulties arise when trying to develop a spectral processing framework. The main difficulty is to define a spectral function basis on a domain which is a 2D (resp. 3D for moving objects) manifold embedded in 3D (resp. 4D) space and thus has an arbitrary topology and a possibly complicated geometry.

3.4. Surface Deformation

Recovering the temporal evolution of a deformable surface is a fundamental task in computer vision, with a large variety of applications ranging from the motion capture of articulated shapes, such as human bodies, to the deformation of complex surfaces such as clothes. Methods that solve for this problem usually infer surface evolutions from motion or geometric cues. This information can be provided by motion capture systems or one of the numerous available static 3D acquisition modalities. In this inference, methods are faced with the challenging estimation of the time-consistent deformation of a surface from cues that can be sparse and noisy. Such an estimation is an ill posed problem that requires prior knowledge on the deformation to be introduced in order to limit the range of possible solutions.

3.5. Manifold Learning

The goal of motion analysis is to understand the movement in terms of movement coordination and corresponding neuromotor and biomechanical principles. Most existing tools for motion analysis consider as input rotational parameters obtained through an articulated body model, e.g. a skeleton; such model being tracked using markers or estimated from shape information. Articulated motion is then traditionally represented by trajectories of rotational data, each rotation in space being associated to the orientation of one limb segment in the body model. This offers a high dimensional parameterization of all possible poses. Typically, using a standard set of articulated segments for a 3D skeleton, this parameterization offers a number of degrees of freedom (DOF) that ranges from 30 to 40. However, it is well known that for a given motion performance, the trajectories of these DOF span a much reduced space. Manifold learning techniques on rotational data have proven their relevance to represent various motions into subspaces of high-level parameters. However, rotational data encode motion information only, independently of morphology, thus hiding the influence of shapes over motion parameters. One of the objectives is to investigate how motions of human and animal bodies, i.e. dense surface data, span manifolds in higher dimensional spaces and how these manifolds can be characterized. The main motivation is to propose morpho-dynamic indices of motion that account for both shape and motion. Dimensionality reduction will be applied on these data and used to characterize the manifolds associated to human motions. To this purpose, the raw mesh structure cannot be statistically processed directly and appropriate features extraction as well as innovative multidimensional methods must be investigated.

PERCEPTION Team

3. Scientific Foundations

3.1. The geometry of multiple images

Computer vision requires models that describe the image creation process. An important part (besides e.g. radiometric effects), concerns the geometrical relations between the scene, cameras and the captured images, commonly subsumed under the term "multi-view geometry". This describes how a scene is projected onto an image, and how different images of the same scene are related to one another. Many concepts are developed and expressed using the tool of projective geometry. As for numerical estimation, e.g. structure and motion calculations, geometric concepts are expressed algebraically. Geometric relations between different views can for example be represented by so-called matching tensors (fundamental matrix, trifocal tensors, ...). These tools and others allow to devise the theory and algorithms for the general task of computing scene structure and camera motion, and especially how to perform this task using various kinds of geometrical information: matches of geometrical primitives in different images, constraints on the structure of the scene or on the intrinsic characteristics or the motion of cameras, etc.

3.2. The photometry component

In addition to the geometry (of scene and cameras), the way an image looks like depends on many factors, including illumination, and reflectance properties of objects. The reflectance, or "appearance", is the set of laws and properties which govern the radiance of the surfaces . This last component makes the connections between the others. Often, the "appearance" of objects is modeled in image space, e.g. by fitting statistical models, texture models, deformable appearance models (...) to a set of images, or by simply adopting images as texture maps.

Image-based modelling of 3D shape, appearance, and illumination is based on prior information and measures for the coherence between acquired images (data), and acquired images and those predicted by the estimated model. This may also include the aspect of temporal coherence, which becomes important if scenes with deformable or articulated objects are considered.

Taking into account changes in image appearance of objects is important for many computer vision tasks since they significantly affect the performances of the algorithms. In particular, this is crucial for feature extraction, feature matching/tracking, object tracking, 3D modelling, object recognition etc.

3.3. Shape Acquisition

Recovering shapes from images is a fundamental task in computer vision. Applications are numerous and include, in particular, 3D modeling applications and mixed reality applications where real shapes are mixed with virtual environments. The problem faced here is to recover shape information such as surfaces, point positions, or differential properties from image information. A tremendous research effort has been made in the past to solve this problem and a number of partial solutions had been proposed. However, a fundamental issue still to be addressed is the recovery of full shape information over time sequences. The main difficulties are precision, robustness of computed shapes as well as consistency of these shapes over time. An additional difficulty raised by real-time applications is complexity. Such applications are today feasible but often require powerful computation units such as PC clusters. Thus, significant efforts must also be devoted to switch from traditional single-PC units to modern computation architectures.

3.4. Motion Analysis

The perception of motion is one of the major goals in computer vision with a wide range of promising applications. A prerequisite for motion analysis is motion modelling. Motion models span from rigid motion to complex articulated and/or deformable motion. Deformable objects form an interesting case because the models are closely related to the underlying physical phenomena. In the recent past, robust methods were developed for analysing rigid motion. This can be done either in image space or in 3D space. Image-space analysis is appealing and it requires sophisticated non-linear minimization methods and a probabilistic framework. An intrinsic difficulty with methods based on 2D data is the ambiguity of associating a multiple degree of freedom 3D model with image contours, texture and optical flow. Methods using 3D data are more relevant with respect to our recent research investigations. 3D data are produced using stereo or a multiple-camera setup. These data (surface patches, meshes, voxels, etc.) are matched against an articulated object model (based on cylindrical parts, implicit surfaces, conical parts, and so forth). The matching is carried out within a probabilistic framework (pair-wise registration, unsupervised learning, maximum likelihood with missing data).

Challenging problems are the detection and segmentation of multiple moving objects and of complex articulated objects, such as human-body motion, body-part motion, etc. It is crucial to be able to detect motion cues and to interpret them in terms of moving parts, independently of a prior model. Another difficult problem is to track articulated motion over time and to estimate the motions associated with each individual degree of freedom.

3.5. Multiple-camera acquisition of visual data

Modern computer vision techniques and applications require the deployment of a large number of cameras linked to a powerful multi-PC computing platform. Therefore, such a system must fulfill the following requirements: The cameras must be synchronized up to the millisecond, the bandwidth associated with image transfer (from the sensor to the computer memory) must be large enough to allow the transmission of uncompressed images at video rates, and the computing units must be able to dynamically store the data and/to process them in real-time.

Current camera acquisition systems are all-digital ones. They are based on standard network communication protocols such as the IEEE 1394. Recent systems involve as well depth cameras that produce depth images, i.e. a depth information at each pixel. Popular technologies for this purpose include the Time of Flight Cameras (TOF cam) and structured light cameras, as in the very recent Microsoft's Kinect device.

3.6. Auditory and audio-visual scene analysis

For the last two years, PERCEPTION has started to investigate a new research topic, namely the analysis of auditory information and the fusion between auditory and visual data. In particular we are interested in analyzing the acoustic layout of a scene (how many sound sources are out there and where are they located? what is the semantic content of each auditory signal?) For that purpose we use microphones that are mounted onto a human-like head. This allows the extraction of several kinds of auditory cues, either based on the time difference of arrival or based on the fact that the head and the ears modify the spectral properties of the sounds perceived with the left and right microphones. Both the temporal and spectral binaural cues can be used to locate the most proeminent sound sources, and to separate the perceived signal into several sources. This is however an extremely difficult task because of the inherent ambiguity due resemblance of signals, and of the presence of acoustic noise and reverberations. The combination of visual and auditory data allows to solve the localization and separation tasks in a more robust way, provided that the two stimuli are available. One interesting yet unexplored topic is the development of hearing for robots, such as the role of head and body motions in the perception of sounds.

PRIMA Project-Team

3. Scientific Foundations

3.1. Context Aware Smart Spaces

Situation Models for Context Aware Systems and Services

3.1.1. Summary

Over the last few years, the PRIMA group has pioneered the use of context aware observation of human activity in order to provide non-disruptive services. In particular, we have developed a conceptual framework for observing and modeling human activity, including human-to-human interaction, in terms of situations.

Encoding activity in situation models provides a formal representation for building systems that observe and understand human activity. Such models provide scripts of activities that tell a system what actions to expect from each individual and the appropriate behavior for the system. A situation model acts as a non-linear script for interpreting the current actions of humans, and predicting the corresponding appropriate and inappropriate actions for services. This framework organizes the observation of interaction using a hierarchy of concepts: scenario, situation, role, action and entity. Situations are organized into networks, with transition probabilities, so that possible next situations may be predicted from the current situation.

Current technology allows us to handcraft real-time systems for a specific services. The current hard challenge is to create a technology to automatically learn and adapt situation models with minimal or no disruption of human activity. An important current problem for the PRIMA group is the adaptation of Machine Learning techniques for learning situation models for describing the context of human activity.

3.1.2. Detailed Description

Context Aware Systems and Services require a model for how humans think and interact with each other and their environment. Relevant theories may be found in the field of cognitive science. Since the 1980's, Philippe Johnson-Laird and his colleagues have developed an extensive theoretical framework for human mental models [52], [53]. Johnson Laird's "situation models", provide a simple and elegant framework for predicting and explaining human abilities for spatial reasoning, game playing strategies, understanding spoken narration, understanding text and literature, social interaction and controlling behavior. While these theories are primarily used to provide models of human cognitive abilities, they are easily implemented in programmable systems [37], [36].

In Johnson-Laird's Situation Models, a situation is defined as a configuration of relations over entities. Relations are formalized as N-ary predicates such as beside or above. Entities are objects, actors, or phenomena that can be reliably observed by a perceptual system. Situation models provide a structure for organizing assemblies of entities and relations into a network of situations. For cognitive scientists, such models provide a tool to explain and predict the abilities and limitations of human perception. For machine perception systems, situation models provide the foundation for assimilation, prediction and control of perception. A situation model identifies the entities and relations that are relevant to a context, allowing the perception system to focus limited computing and sensing resources. The situation model can provide default information about the identities of entities and the configuration of relations, allowing a system to continue to operate when perception systems fail or become unreliable. The network of situations provides a mechanism to predict possible changes in entities or their relations. Finally, the situation model provides an interface between perception and human centered systems and services. On the one hand, changes in situations can provide events that drive service behavior. At the same time, the situation model can provide a default description of the environment that allows human-centered services to operate asynchronously from perceptual systems.

We have developed situation models based on the notion of a script. A theatrical script provides more than dialog for actors. A script establishes abstract characters that provide actors with a space of activity for expression of emotion. It establishes a scene within which directors can layout a stage and place characters. Situation models are based on the same principle.

A script describes an activity in terms of a scene occupied by a set of actors and props. Each actor plays a role, thus defining a set of actions, including dialog, movement and emotional expressions. An audience understands the theatrical play by recognizing the roles played by characters. In a similar manner, a user service uses the situation model to understand the actions of users. However, a theatrical script is organised as a linear sequence of scenes, while human activity involves alternatives. In our approach, the situation model is not a linear sequence, but a network of possible situations, modeled as a directed graph.

Situation models are defined using roles and relations. A role is an abstract agent or object that enables an action or activity. Entities are bound to roles based on an acceptance test. This acceptance test can be seen as a form of discriminative recognition.

There is no generic algorithm capable of robustly recognizing situations from perceptual events coming from sensors. Various approaches have been explored and evaluated. Their performance is very problem and environment dependent. In order to be able to use several approaches inside the same application, it is necessary to clearly separate the specification of context (scenario) and the implementation of the program that recognizes it, using a Model Driven Engineering approach. The transformation between a specification and its implementation must be as automatic as possible. We have explored three implementation models :

Synchronized petri net. The Petri Net structure implements the temporal constraints of the initial context model (Allen operators). The synchronisation controls the Petri Net evolution based on roles and relations perception. This approach has been used for the Context Aware Video Acquisition application (more details at the end of this section).

Fuzzy Petri Nets. The Fuzzy Petri Net naturally expresses the smooth changes of activity states (situations) from one state to another with gradual and continuous membership function. Each fuzzy situation recognition is interpreted as a new proof of the recognition of the corresponding context. Proofs are then combined using fuzzy integrals. This approach has been used to label videos with a set of predefined scenarios (context).

Hidden Markov Model. This probabilistic implementation of the situation model integrates uncertainty values that can both refer to confidence values for events and to a less rigid representation of situations and situations transitions. This approach has been used to detect interaction groups (in a group of meeting participants, who is interacting with whom and thus which interaction groups are formed)

Currently situation models are constructed by hand. Our current challenge is to provide a technology by which situation models may be adapted and extended by explicit and implicit interaction with the user. An important aspect of taking services to the real world is an ability to adapt and extend service behaviour to accommodate individual preferences and interaction styles. Our approach is to adapt and extend an explicit model of user activity. While such adaptation requires feedback from users, it must avoid or at least minimize disruption. We are curently exploring reinforcement learning approaches to solve this problem.

With a reinforcement learning approach, the system is rewarded and punished by user reactions to system behaviors. A simplified stereotypic interaction model assures a initial behavior. This prototypical model is adapted to each particular user in a way that maximizes its satisfaction. To minimize distraction, we are using an indirect reinforcement learning approach, in which user actions and consequences are logged, and this log is periodically used for off-line reinforcement learning to adapt and refine the context model.

Adaptations to the context model can result in changes in system behaviour. If unexpected, such changes may be disturbing for the end users. To keep user's confidence, the learned system must be able to explain its actions. We are currently exploring methods that would allow a system to explain its model of interaction. Such explanation is made possible by explicit describing context using situation models.

The PRIMA group has refined its approach to context aware observation in the development of a process for real time production of a synchronized audio-visual stream based using multiple cameras, microphones and other information sources to observe meetings and lectures. This "context aware video acquisition system" is an automatic recording system that encompasses the roles of both the camera-man and the director. The system determines the target for each camera, and selects the most appropriate camera and microphone to record the current activity at each instant of time. Determining the most appropriate camera and microphone requires a model of activities of the actors, and an understanding of the video composition rules. The model of the activities of the actors is provided by a "situation model" as described above.

In collaboration with France Telecom, we have adapted this technology to observing social activity in domestic environments. Our goal is to demonstrate new forms of services for assisted living to provide non-intrusive access to care as well to enhance informal contact with friends and family.

3.2. Service Oriented Architectures for Intelligent Environments

Software Architecture, Service Oriented Computing, Service Composition, Service Factories, Semantic Description of Functionalities

Intelligent environments are at the confluence of multiple domains of expertise. Experimenting within intelligent environments requires combining techniques for robust, autonomous perception with methods for modeling and recognition of human activity within an inherently dynamic environment. Major software engineering and architecture challenges include accomodation of a heterogeneous of devices and software, and dynamically adapting to changes human activity as well as operating conditions.

The PRIMA project explores software architectures that allow systems to be adapt to individual user preferences. Interoperability and reuse of system components is fundamental for such systems. Adopting a shared, common Service Oriented Architecture (SOA) architecture has allowed specialists from a variety of subfields to work together to build novel forms of systems and services.

In a service oriented architecture, each hardware or software component is exposed to the others as a “service”. A service exposes its functionality through a well defined interface that abstracts all the implementation details and that is usually available through the network.

The most commonly known example of a service oriented architecture are the Web Services technologies that are based on web standards such as HTTP and XML. Semantic Web Services proposes to use knowledge representation methods such as ontologies to give some semantic to services functionalities. Semantic description of services makes it possible to improve the interoperability between services designed by different persons or vendors.

Taken out of the box, most SOA implementations have some “defects” preventing their adoption. Web services, due to their name, are perceived as being only for the “web” and also as having a notable performance overhead. Other implementations such as various propositions around the Java virtual machine, often requires to use a particular programming language or are not distributed. Intelligent environments involves many specialist and a hard constraint on the programming language can be a real barrier to SOA adoption.

The PRIMA project has developed OMiSCID, a middleware for service oriented architectures that addresses the particular problematics of intelligent environments. OMiSCID has emerged as an effective tool for unifying access to functionalities provided from the lowest abstraction level components (camera image acquisition, image processing) to abstract services such (activity modeling, personal assistant). OMiSCID has facilitated cooperation by experts from within the PRIMA project as well as in projects with external partners.

Experiments with semantic service description and spontaneous service composition are conducted around the OMiSCID middleware. In these experiments, attention is paid to usability. A dedicated language has been designed to allow developers to describe the functionalities that their services provide. This language aims at simplifying existing semantic web services technologies to make them usable by a normal developer (i.e. that is not specialized in the semantic web). This language is named the User-oriented Functionality Composition Language (UFCL).

UFCL allows developers to specify three types of knowledge about services:

The knowledge that a service exposes a functionality like a “Timer” functionality for a service emitting message at a regular frequency.

The knowledge that a kind of functionality can be converted to another one. For example, a “Metronome” functionality issued from a music centered application can be seen as a “Timer” functionality.

The knowledge that a particular service is a factory and can instantiate other services on demand. A TimerFactory can for example start a new service with a “Timer” functionality with any desired frequency. Factories greatly helps in the deployment of service based applications. UFCL factories can also express the fact that they can compose existing functionalities to provide another one.

To bring the UFCL descriptions provided by the developers to life, a runtime has been designed to enable reasoning about what functionalities are available, what functionalities can be transformed to another one and what functionalities could be obtained by asking factories. The service looking for a particular functionality has just to express its need in term of functionalities and properties (e.g. a “Timer” with a frequency of 2Hz) and the runtime automates everything else: gathering of UFCL descriptions exposed by all running services, compilation of these descriptions to some rules in a rule-based system, reasoning and creation of a plan to obtained the desired functionality, and potentially invoking service factories to start the missing services.

3.3. Robust view-invariant Computer Vision

Local Appearance, Affine Invariance, Receptive Fields

3.3.1. Summary

A long-term grand challenge in computer vision has been to develop a descriptor for image information that can be reliably used for a wide variety of computer vision tasks. Such a descriptor must capture the information in an image in a manner that is robust to changes the relative position of the camera as well as the position, pattern and spectrum of illumination.

Members of PRIMA have a long history of innovation in this area, with important results in the area of multi-resolution pyramids, scale invariant image description, appearance based object recognition and receptive field histograms published over the last 20 years. The group has most recently developed a new approach that extends scale invariant feature points for the description of elongated objects using scale invariant ridges. PRIMA has worked with ST Microelectronics to embed its multi-resolution receptive field algorithms into low-cost mobile imaging devices for video communications and mobile computing applications.

3.3.2. Detailed Description

The visual appearance of a neighbourhood can be described by a local Taylor series [54]. The coefficients of this series constitute a feature vector that compactly represents the neighbourhood appearance for indexing and matching. The set of possible local image neighbourhoods that project to the same feature vector are referred to as the "Local Jet". A key problem in computing the local jet is determining the scale at which to evaluate the image derivatives.

Lindeberg [56] has described scale invariant features based on profiles of Gaussian derivatives across scales. In particular, the profile of the Laplacian, evaluated over a range of scales at an image point, provides a local description that is "equi-variant” to changes in scale. Equi-variance means that the feature vector translates exactly with scale and can thus be used to track, index, match and recognize structures in the presence of changes in scale.

A receptive field is a local function defined over a region of an image [62]. We employ a set of receptive fields based on derivatives of the Gaussian functions as a basis for describing the local appearance. These functions resemble the receptive fields observed in the visual cortex of mammals. These receptive fields are applied to color images in which we have separated the chrominance and luminance components. Such functions are easily normalized to an intrinsic scale using the maximum of the Laplacian [56], and normalized in orientation using direction of the first derivatives [62].

The local maxima in x and y and scale of the product of a Laplacian operator with the image at a fixed position provides a "Natural interest point" [57]. Such natural interest points are salient points that may be robustly detected and used for matching. A problem with this approach is that the computational cost of determining intrinsic scale at each image position can potentially make real-time implementation unfeasible.

A vector of scale and orientation normalized Gaussian derivatives provides a characteristic vector for matching and indexing. The oriented Gaussian derivatives can easily be synthesized using the "steerability property" [47] of Gaussian derivatives. The problem is to determine the appropriate orientation. In earlier work by PRIMA members Colin de Verdiere [34], Schiele [62] and Hall [50], proposed normalising the local jet independently at each pixel to the direction of the first derivatives calculated at the intrinsic scale. This has provided promising results for many view invariant image recognition tasks as described in the next section.

Color is a powerful discriminator for object recognition. Color images are commonly acquired in the Cartesian color space, RGB. The RGB color space has certain advantages for image acquisition, but is not the most appropriate space for recognizing objects or describing their shape. An alternative is to compute a Cartesian representation for chrominance, using differences of R, G and B. Such differences yield color opponent receptive fields resembling those found in biological visual systems.

Our work in this area uses a family of steerable color opponent filters developed by Daniela Hall [50]. These filters transform an (R,G,B), into a cartesian representation for luminance and chrominance (L,C1,C2). Chromatic Gaussian receptive fields are computed by applying the Gaussian derivatives independently to each of the three components, (L, C1, C2). The components C1 and C2 encodes the chromatic information in a Cartesian representation, while L is the luminance direction. Chromatic Gaussian receptive fields are computed by applying the Gaussian derivatives independently to each of the three components, (L, C1, C2). Permutations of RGB lead to different opponent color spaces. The choice of the most appropriate space depends on the chromatic composition of the scene. An example of a second order steerable chromatic basis is the set of color opponent filters shown in figure 1 .

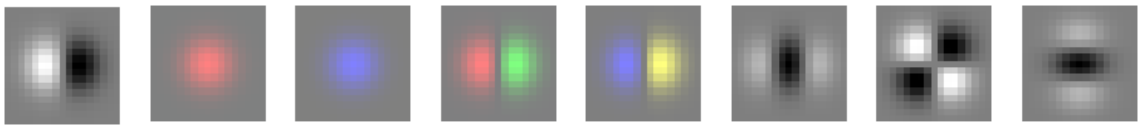


Figure 1. Chromatic Gaussian Receptive Fields ($G_x^L, G^{C_1}, G^{C_2}, G_x^{C_1}, G_x^{C_2}, G_{xx}^L, G_{xy}^L, G_{yy}^L$).

Key results in this area include

Fast, video rate, calculation of scale and orientation for image description with normalized chromatic receptive fields [37].

Real time indexing and recognition using a novel indexing tree to represent multi-dimensional receptive field histograms [60].

Robust visual features for face tracking [49], [48].

Affine invariant detection and tracking using natural interest lines [63].

Direct computation of time to collision over the entire visual field using rate of change of intrinsic scale [58].

We have achieved video rate calculation of scale and orientation normalised Gaussian receptive fields using an O(N) pyramid algorithm [37]. This algorithm has been used to propose an embedded system that provides real time detection and recognition of faces and objects in mobile computing devices.

Applications have been demonstrated for detection, tracking and recognition at video rates. This method has been used in the MinImage project to provide real time detection, tracking, and identification of faces. It has also been used to provide techniques for estimating age and gender of people from their faces

3.4. Perception for Social Interaction

Affective Computing, Perception for social interaction.

Current research on perception for interaction primarily focuses on recognition and communication of linguistic signals. However, most human-to-human interaction is non-verbal and highly dependent on social context. A technology for natural interaction will require abilities to perceive and assimilate non-verbal social signals, to understand and predict social situations, and to acquire and develop social interaction skills.

The overall goal of this research program is to provide the scientific and technological foundations for systems that observe and interact with people in a polite, socially appropriate manner. We address these objectives with research activities in three interrelated areas:

Multimodal perception for social interactions.

Learning models for context aware social interaction, and

Context aware systems and services.

Our approach to each of these areas is to draw on models and theories from the cognitive and social sciences, human factors, and software architectures to develop new theories and models for computer vision and multimodal interaction. Results will be developed, demonstrated and evaluated through the construction of systems and services for polite, socially aware interaction in the context of smart habitats.

3.4.1. Detailed Description

First part of our work on perception for social interaction has concentrated on measuring the physiological parameters of Valence, Arousal and Dominance using visual observation form environmental sensors as well as observation of facial expressions.

People express and feel emotions with their face. Because the face is the both externally visible and the seat of emotional expression, facial expression of emotion plays a central role in social interaction between humans. Thus visual recognition of emotions from facial expressions is a core enabling technology for any effort to adapt ICT for social interaction.

Constructing a technology for automatic visual recognition of emotions requires solutions to a number of hard challenges. Emotions are expressed by coordinated temporal activations of 21 different facial muscles assisted by a number of additional muscles. Activations of these muscles are visible through subtle deformations in the surface structure of the face. Unfortunately, this facial structure can be masked by facial markings, makeup, facial hair, glasses and other obstructions. The exact facial geometry, as well as the coordinated expression of muscles is unique to each individual. In additions, these deformations must be observed and measured under a large variety of illumination conditions as well as a variety of observation angles. Thus the visual recognition of emotions from facial expression remains a challenging open problem in computer vision.

Despite the difficulty of this challenge, important progress has been made in the area of automatic recognition of emotions from face expressions. The systematic cataloging of facial muscle groups as facial action units by Ekman [45] has let a number of research groups to develop libraries of techniques for recognizing the elements of the FACS coding system [33]. Unfortunately, experiments with that system have revealed that the system is very sensitive to both illumination and viewing conditions, as well as the difficulty in interpreting the resulting activation levels as emotions. In particular, this approach requires a high-resolution image with a high signal-to-noise ratio obtained under strong ambient illumination. Such restrictions are not compatible with the mobile imaging system used on tablet computers and mobile phones that are the target of this effort.

As an alternative to detecting activation of facial action units by tracking individual face muscles, we propose to measure physiological parameters that underlie emotions with a global approach. Most human emotions can be expressed as trajectories in a three dimensional space whose features are the physiological parameters of Pleasure-Displeasure, Arousal-Passivity and Dominance-Submission. These three physiological parameters can be measured in a variety of manners including on-body accelerometers, prosody, heart-rate, head movement and global face expression.

In our work, we address the recognition of social behaviors multimodal information. These are unconscious inmate cognitive processes that are vital to human communication and interaction. Recognition of social behaviors enables anticipation and improves the quality of interaction between humans. Among social behaviors, we have focused on engagement, the expression of intention for interaction. During the engagement phase, many non-verbal signals are used to communicate the intention to engage to the partner [65]. These include posture, gaze, spatial information, gestures, and vocal cues.

For example , within the context of frail or elderly people at home, a companion robot must also be able to detect the engagement of humans in order to adapt their responses during interaction with humans to increase their acceptability. Classical approaches for engagement with robots use spatial information such as human position and speed, human-robot distance and the angle of arrival. Our believe that, while such uni-modal methods may be suitable for static display [66] or robots in wide space area [55] they are not sufficient for home environments. In an apartment, relative spatial information of people and robot are not as discriminative as in an open space. Passing by the robot in a corridor should not lead to an engagement detection, and possible socially inappropriate behavior by the robot.

In our experiments, we use a kompai robot from Robosoft [32]. As an alternative to wearable physiological sensors (such as pulse bracelet Cardiocam, etc.) we inegrate multimodal features using a Kinect sensor (see figure 5). In addition of the spatial cues from the laser telemeter, one can use new multimodal features based on persons and skeletons tracking, sound localization, etc. Some of these new features are inspired from results in cognitive science domain [61].

Our multimodal approach has been confronted to a robot centered dataset for multimodal social signal processing recorded in a home-like environment [24]. The evaluation on our corpus highlights its robustness and validates use of such technique in real environment. Experimental validation shows that the use of multimodal sensors gives better results than only spatial features (50% of error reduction). Our experimentations also confirm results from [61]: relative shoulder rotation, speed and facing visage are among crucial features for engagement detection.

3.5. End Users control over Smart Environment

Missing keywords.

Ubiquitous computing promises unprecedented empowerment from the flexible and robust combination of software services with the physical world. Software researchers assimilate this promise as system autonomy where users are conveniently kept out of the loop. Their hypothesis is that services, such as music playback and calendars, are developed by service providers and pre-assembled by software designers to form new service frontends. Their scientific challenge is then to develop secure, multiscale, multi-layered, virtualized infrastructures that guarantee service front-end continuity. Although service continuity is desirable in many circumstances, end users, with this interpretation of ubiquitous computing, are doomed to behave as mere consumers, just like with conventional desktop computing.

Another interpretation of the promises of ubiquitous computing, is the empowerment of end users with tools that allow them to create and reshape their own interactive spaces. Our hypothesis is that end users are willing to shape their own interactive spaces by coupling smart artifacts, building imaginative new functionalities that were not anticipated by system designers. A number of tools and techniques have been developed to support this view such as CAMP [64] or iCAP [44].

We adopt a End-User Programming (EUP) approach to give the control back to the inhabitants. In our vision, smart Homes will be incrementally equiped with sensors, actuators and services by inhabitants themselves. Our research programm therefore focus on tools and languages to enable inhabitants in activities related to EUP for Smart Homes :

Installation and maintenance of devices and services. This may imply having facilities to attribute names.

Visualizing and controling of the Smart Habitat.

Programming and testing. This imply one or more programming languages and programming environment which could rely on the previous point. The programming language is especially important. Indeed, in the context of the Smart Homes, End-User Programms are most likely to be routines in the sens of [38] than procedure in the sens of traditionnal programming languages.

Detecting and solving conflicts related to contradictory programms or goals.

SIROCCO Project-Team

3. Scientific Foundations

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 modelling, of signal processing, of coding and information theory. However, the objective of better exploiting the Human Visual System (HVS) properties in the above goals also pertains to the areas of perceptual modelling and cognitive science. Some of the proposed research axes are also based on scientific foundations of computer vision (e.g. multi-view modelling 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. Parameter estimation and inference

Bayesian estimation, Expectation-Maximization, stochastic modelling

Parameter estimation is at the core of the processing tools studied and developed in the team. Applications range from the prediction of missing data or future data, to extracting some information about the data in order to perform efficient compression. More precisely, the data are assumed to be generated by a given stochastic data model, which is partially known. The set of possible models translates the a priori knowledge we have on the data and the best model has to be selected in this set. When the set of models or equivalently the set of probability laws is indexed by a parameter (scalar or vectorial), the model is said parametric and the model selection resorts to estimating the parameter. Estimation algorithms are therefore widely used at the encoder in order to analyze the data. In order to achieve high compression rates, the parameters are usually not sent and the decoder has to jointly select the model (i.e. estimate the parameters) and extract the information of interest.

3.3. Data Dimensionality Reduction

manifolds, locally linear embedding, non-negative matrix factorization, principal component analysis

A fundamental problem in many data processing tasks (compression, classification, indexing) is to find a suitable representation of the data. It often aims at reducing the dimensionality of the input data so that tractable processing methods can then be applied. Well-known methods for data dimensionality reduction include the principal component analysis (PCA) and independent component analysis (ICA). The methodologies which will be central to several proposed research problems will instead be based on sparse representations, on locally linear embedding (LLE) and on the "non negative matrix factorization" (NMF) framework.

The objective of *sparse representations* is to find a sparse approximation of a given input data. In theory, given $A \in \mathbb{R}^{m \times n}$, $m < n$, and $\mathbf{b} \in \mathbb{R}^m$ with $m << n$ and A is of full 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 x, i.e. the number of non-zero components in z. There exist many solutions x to $Ax = b$. The problem is to find the sparsest, the one for which x has the fewest non zero 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 norm p is usually 2, but could be 1 or ∞ as well. Except for the exhaustive combinatorial approach, there is no known method to find the exact solution under general conditions on the dictionary A. Searching for this sparsest representation is hence unfeasible and both problems are computationally intractable. Pursuit algorithms have been introduced as heuristic methods which aim at finding approximate solutions to the above problem with tractable complexity.

Non negative matrix factorization (NMF) is a non-negative approximate data representation [3]. NMF aims at finding an approximate factorization of a non-negative input data matrix V into non-negative matrices W and H, where the columns of W can be seen as *basis vectors* and those of H as coefficients of the linear approximation of the input data. Unlike other linear representations like principal component analysis (PCA) and independent component analysis (ICA), the non-negativity constraint makes the representation purely additive. Classical data representation methods like PCA or Vector Quantization (VQ) can be placed in an NMF framework, the differences arising from different constraints being placed on the W and H matrices. In VQ, each column of H is constrained to be unary with only one non-zero coefficient which is equal to 1. In PCA, the columns of W are constrained to be orthonormal and the rows of H to be orthogonal to each other. These methods of data-dependent dimensionality reduction will be at the core of our visual data analysis and compression activities.

3.4. Perceptual Modelling

Saliency, visual attention, cognition

The human visual system (HVS) is not able to process all visual information of our visual field at once. To cope with this problem, our visual system must filter out the irrelevant information and reduce redundant information. This feature of our visual system is driven by a selective sensing and analysis process. For instance, it is well known that the greatest visual acuity is provided by the fovea (center of the retina). Beyond this area, the acuity drops down with the eccentricity. Another example concerns the light that impinges on our retina. Only the visible light spectrum lying between 380 nm (violet) and 760 nm (red) is processed. To conclude on the selective sensing, it is important to mention that our sensitivity depends on a number of factors such as the spatial frequency, the orientation or the depth. These properties are modeled by a sensitivity function such as the Contrast Sensitivity Function (CSF).

Our capacity of analysis is also related to our visual attention. Visual attention which is closely linked to eye movement (note that this attention is called overt while the covert attention does not involve eye movement) allows us to focus our biological resources on a particular area. It can be controlled by both top-down (i.e. goal-directed, intention) and bottom-up (stimulus-driven, data-dependent) sources of information [4]. This detection is also influenced by prior knowledge about the environment of the scene[5]. Implicit assumptions related to Prior knowledge or beliefs form play an important role in our perception (see the example concerning the assumption that light comes from above-left). Our perception results from the combination of prior beliefs with data we gather from the environment. A Bayesian framework is an elegant solution to model these interactions[6]. We define a vector $\overrightarrow{v}_l$ of local measurements (contrast of color, orientation, etc.) and vector $\overrightarrow{v}_c$ of global and contextual features (global features, prior locations, type of the scene, etc.). The salient locations S for a spatial position $\overrightarrow{x}$ are then given by:

$$S(\overrightarrow{x}) = \frac{1}{p(\overrightarrow{v}_l \,|\, \overrightarrow{v}_c)} \times p(s, \overrightarrow{x} \,|\, \overrightarrow{v}_c) \tag{5}$$

The first term represents the bottom-up salience. It is based on a kind of contrast detection, following the assumption that rare image features are more salient than frequent ones. Most of existing computational models of visual attention rely on this term. However, different approaches exist to extract the local visual features as well as the global ones. The second term is the contextual priors. For instance, given a scene, it indicates which parts of the scene are likely the most salient.

[3] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization", Nature 401, 6755, (Oct. 1999), pp. 788-791.

[4] L. Itti and C. Koch, "Computational Modelling of Visual Attention" , Nature Reviews Neuroscience, Vol. 2, No. 3, pp. 194-203, 2001.

[5] J. Henderson, "Regarding scenes", Directions in Psychological Science, vol. 16, pp. 219-222, 2007.

[6] L. Zhang, M. Tong, T. Marks, H. Shan, H. and G.W. Cottrell, "SUN: a Bayesian framework for saliency using natural statistics",Journal of Vision, vol. 8, pp. 1-20, 2008.

3.5. 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 [7] 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[8] (SW) and Wyner-Ziv [9] (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.

[7] T. M. Cover and J. A. Thomas, Elements of Information Theory, Second Edition, July 2006.

[8] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources." IEEE Transactions on Information Theory, 19(4), pp. 471-480, July 1973.

[9] 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 Team

3. Scientific Foundations

3.1. Introduction

Stars follows three main research directions: perception for activity recognition, semantic activity recognition, and software engineering for activity recognition. **These three research directions are interleaved**: the software architecture direction provides new methodologies for building safe activity recognition systems and the perception and the semantic activity recognition directions provide new activity recognition techniques which are designed and validated for concrete video analytics and healthcare applications. Conversely, these concrete systems raise new software issues that enrich the software engineering research direction.

Transversally, 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: Guillaume Charpiat, François Brémond, Sabine Moisan, Monique Thonnat.

Computer Vision; Cognitive Systems; Learning; Activity Recognition.

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 detections 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 sensors (e.g. Kinect) 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 videosurveillance and several days in healthcare). To guarantee the long term coherence of tracked objects, spatio-temporal reasoning is required. Modelling 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 framework1pt modelling 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.

Controling 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.

3.2.3. Learning shape and motion

Another approach, to improve jointly segmentation and tracking, is to consider videos as 3D volumetric data and to search for trajectories of points that are statistically coherent both spatially and temporally. This point of view enables new kinds of statistical segmentation criteria and ways to learn them.

We are also using the shape statistics developed in [5] for the segmentation of images or videos with shape prior, by learning local segmentation criteria that are suitable for parts of shapes. This unifies patch-based detection methods and active-contour-based segmentation methods in a single framework. These shape statistics can be used also for a fine classification of postures and gestures, in order to extract more precise information from videos for further activity recognition. In particular, the notion of shape dynamics has to be studied.

More generally, to improve segmentation quality and speed, different optimization tools such as graph-cuts can be used, extended or improved.

3.3. Semantic Activity Recognition

Participants: Guillaume Charpiat, François Brémond, Sabine Moisan, Monique Thonnat.

Activity Recognition, Scene Understanding,Computer Vision

3.3.1. 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 analysing 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 two following research axes: high level understanding (to recognize the activities of physical objects based on high level activity models) and learning (how to learn the models needed for activity recognition).

3.3.2. 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 modelling 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. We have also built a language for video event modelling 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 ...).

3.3.3. 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.

3.3.4. 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, they will be detailed in section 3.4 .

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).

3.4. Software Engineering for Activity Recognition

Participants: Sabine Moisan, Annie Ressouche, Jean-Paul Rigault, François Brémond.

Software Engineering, Generic Components, Knowledge-based Systems, Software Component Platform, Object-oriented Frameworks, Software Reuse, Model-driven Engineering

The aim of this research axis is to build general solutions and tools to develop systems dedicated to activity recognition. For this, we rely on state-of-the art Software Engineering practices to ensure both sound design and easy use, providing genericity, modularity, adaptability, reusability, extensibility, dependability, and maintainability.

This research requires theoretical studies combined with validation based on concrete experiments conducted in Stars. We work on the following three research axes: models (adapted to the activity recognition domain), platform architecture (to cope with deployment constraints and run time adaptation), and system verification (to generate dependable systems). For all these tasks we follow state of the art Software Engineering practices and, if needed, we attempt to set up new ones.

3.4.1. Platform Architecture for Activity Recognition

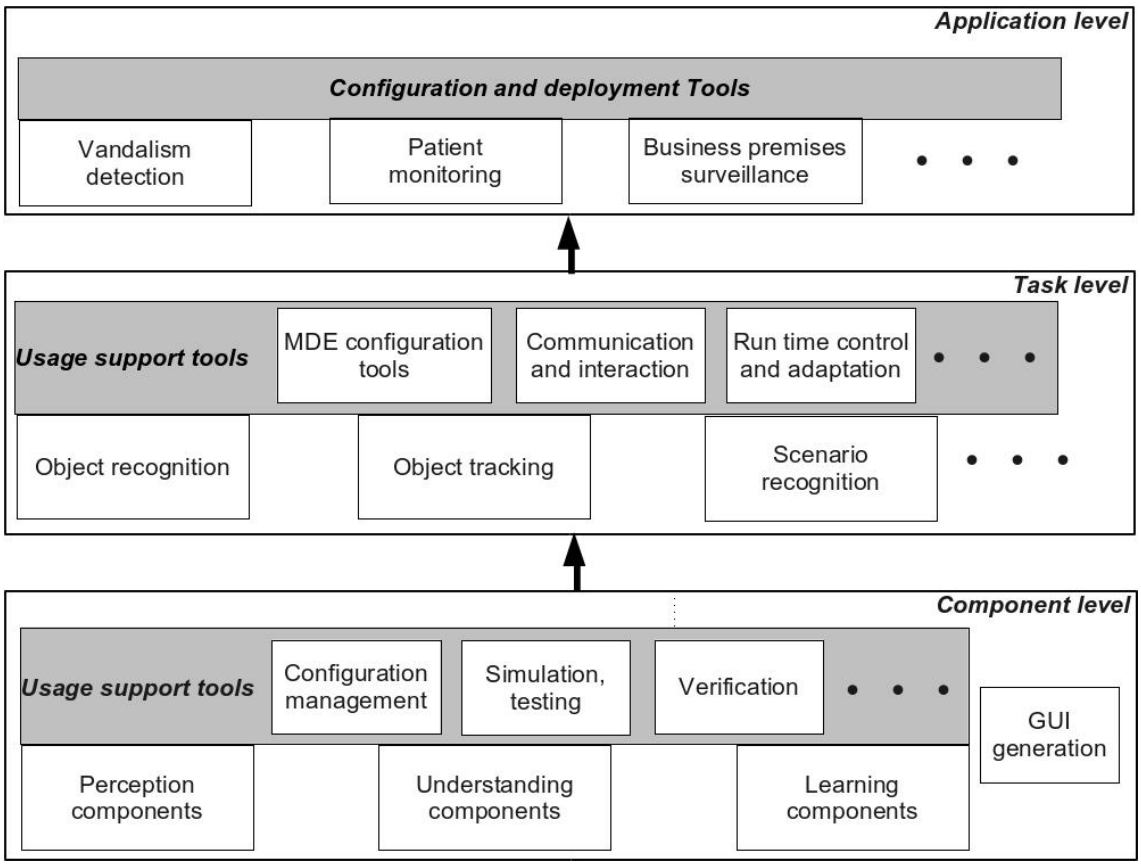


Figure 4. Global Architecture of an Activity Recognition The grey areas contain software engineering support modules whereas the other modules correspond to software components (at Task and Component levels) or to generated systems (at Application level).

In the former project teams Orion and Pulsar, we have developed two platforms, one (VSIP), a library of real-time video understanding modules and another one, LAMA [15], a software platform enabling to design not only knowledge bases, but also inference engines, and additional tools. LAMA offers toolkits to build and to adapt all the software elements that compose a knowledge-based system or a cognitive system.

Figure 4 presents our conceptual vision for the architecture of an activity recognition platform. It consists of three levels:

- The **Component Level**, the lowest one, offers software components providing elementary operations and data for perception, understanding, and learning.

  - Perception components contain algorithms for sensor management, image and signal analysis, image and video processing (segmentation, tracking...), etc.
  - Understanding components provide the building blocks for Knowledge-based Systems: knowledge representation and management, elements for controlling inference engine strategies, etc.
  - Learning components implement different learning strategies, such as Support Vector Machines (SVM), Case-based Learning (CBL), clustering, etc.

  An Activity Recognition system is likely to pick components from these three packages. Hence, tools must be provided to configure (select, assemble), simulate, verify the resulting component combination. Other support tools may help to generate task or application dedicated languages or graphic interfaces.

- The **Task Level**, the middle one, contains executable realizations of individual tasks that will collaborate in a particular final application. Of course, the code of these tasks is built on top of the components from the previous level. We have already identified several of these important tasks: Object Recognition, Tracking, Scenario Recognition... In the future, other tasks will probably enrich this level.

  For these tasks to nicely collaborate, communication and interaction facilities are needed. We shall also add MDE-enhanced tools for configuration and run-time adaptation.

- The **Application Level** integrates several of these tasks to build a system for a particular type of application, e.g., vandalism detection, patient monitoring, aircraft loading/unloading surveillance, etc.. Each system is parametrized to adapt to its local environment (number, type, location of sensors, scene geometry, visual parameters, number of objects of interest...). Thus configuration and deployment facilities are required.

The philosophy of this architecture is to offer at each level a balance between the widest possible genericity and the maximum effective reusability, in particular at the code level.

To cope with real application requirements, we shall also investigate distributed architecture, real time implementation, and user interfaces.

Concerning implementation issues, we shall use when possible existing open standard tools such as NuSMV for model-checking, Eclipse for graphic interfaces or model engineering support, Alloy for constraint representation and SAT solving, etc. Note that, in Figure 4 , some of the boxes can be naturally adapted from SUP existing elements (many perception and understanding components, program supervision, scenario recognition...) whereas others are to be developed, completely or partially (learning components, most support and configuration tools).

3.4.2. Discrete Event Models of Activities

As mentioned in the previous section (3.3) we have started to specify a formal model of scenario dealing with both absolute time and logical time. Our scenario and time models as well as the platform verification tools rely on a formal basis, namely the synchronous paradigm. To recognize scenarios, we consider activity descriptions as synchronous reactive systems and we apply general modelling methods to express scenario behaviour.

Activity recognition systems usually exhibit many safeness issues. From the software engineering point of view we only consider software security. Our previous work on verification and validation has to be pursued; in particular, we need to test its scalability and to develop associated tools. Model-checking is an appealing technique since it can be automatized and helps to produce a code that has been formally proved. Our verification method follows a compositional approach, a well-known way to cope with scalability problems in model-checking.

Moreover, recognizing real scenarios is not a purely deterministic process. Sensor performance, precision of image analysis, scenario descriptions may induce various kinds of uncertainty. While taking into account this uncertainty, we should still keep our model of time deterministic, modular, and formally verifiable. To formally describe probabilistic timed systems, the most popular approach involves probabilistic extension of timed automata. New model checking techniques can be used as verification means, but relying on model checking techniques is not sufficient. Model checking is a powerful tool to prove decidable properties but introducing uncertainty may lead to infinite state or even undecidable properties. Thus model checking validation has to be completed with non exhaustive methods such as abstract interpretation.

3.4.3. Model-Driven Engineering for Configuration and Control and Control of Video Surveillance systems

Model-driven engineering techniques can support the configuration and dynamic adaptation of video surveillance systems designed with our SUP activity recognition platform. The challenge is to cope with the many—functional as well as nonfunctional—causes of variability both in the video application specification and in the concrete SUP implementation. We have used *feature models* to define two models: a generic model of video surveillance applications and a model of configuration for SUP components and chains. Both of them express variability factors. Ultimately, we wish to automatically generate a SUP component assembly from an application specification, using models to represent transformations [54]. Our models are enriched with intra- and inter-models constraints. Inter-models constraints specify models to represent transformations. Feature models are appropriate to describe variants; they are simple enough for video surveillance experts to express their requirements. Yet, they are powerful enough to be liable to static analysis [70]. In particular, the constraints can be analysed as a SAT problem.

An additional challenge is to manage the possible run-time changes of implementation due to context variations (e.g., lighting conditions, changes in the reference scene, etc.). Video surveillance systems have to dynamically adapt to a changing environment. The use of models at run-time is a solution. We are defining adaptation rules corresponding to the dependency constraints between specification elements in one model and software variants in the other [51], [80], [72].

TEXMEX Project-Team

3. Scientific Foundations

3.1. Image description

In most contexts where images are to be compared, a direct comparison is impossible. Images are compressed in different formats, most formats are error-prone, images are re-sized, cropped, etc. The solution consists in computing descriptors, which are invariant to these transformations.

The first description methods associate a unique global descriptor with each image, *e.g.*, a color histogram or correlogram, a texture descriptor. Such descriptors are easy to compute and use, but they usually fail to handle cropping and cannot be used for object recognition. The most successful approach to address a large class of transformations relies on the use of local descriptors, extracted on regions of interest detected by a detector, for instance the Harris detector [82] or the Difference of Gaussian method proposed by David Lowe [84].

The detectors select a square, circular or elliptic region that is described in turn by a patch descriptor, usually referred to as a local descriptor. The most established description method, namely the SIFT descriptor [84], was shown robust to geometric and photometric transforms. Each local SIFT descriptor captures the information provided by the gradient directions and intensities in the region of interest in each region of a 4×4 grid, thereby taking into account the spatial organization of the gradient in a region. As a matter of fact, the SIFT descriptor has became a standard for image and video description.

Local descriptors can be used in many applications: image comparison for object recognition, image copy detection, detection of repeats in television streams, etc. While they are very reliable, local descriptors are not without problems. As many descriptors can be computed for a single image, a collection of one million images generates in the order of a billion descriptors. That is why specific indexing techniques are required. The problem of taking full advantage of these strong descriptors on a large scale is still an open and active problem. Most of the recent techniques consists in computing a global descriptor from local ones, such as proposed in the so-called bag-of-visual-word approach [89]. Recently, global description computed from local descriptors has been shown successful in breaking the complexity problem. We are active in designing methods that aggregate local descriptors into a single vector representation without loosing too much of the discriminative power of the descriptors.

3.2. Corpus-based text description and machine learning

Our work on textual material (textual documents, transcriptions of speech documents, captions in images or videos, etc.) is characterized by a chiefly corpus-based approach, as opposed to an introspective one. A corpus is for us a huge collection of textual documents, gathered or used for a precise objective. We thus exploit specialized (abstracts of biomedical articles, computer science texts, etc.) or non specialized (newspapers, broadcast news, etc.) collections for our various studies. In TEXMEX, according to our applications, different kinds of knowledge can be extracted from the textual material. For example, we automatically extract terms characteristic of each successive topic in a corpus with no a priori knowledge; we produce representations for documents in an indexing perspective [88]; we acquire lexical resources from the collections (morphological families, semantic relations, translation equivalences, etc.) in order to better grasp relations between segments of texts in which a same idea is expressed with different terms or in different languages...

In the domain of the corpus-based text processing, many researches have been undergone in the last decade. While most of them are essentially based on statistical methods, symbolic approaches also present a growing interest [78]. For our various problems involving language processing, we use both approaches, making the most of existing machine learning techniques or proposing new ones. Relying on advantages of both methods, we aim at developing machine learning solutions that are automatic and generic enough to make it possible to extract, from a corpus, the kind of elements required by a given task.

3.3. Stochastic models for multimodal analysis

Describing multimedia documents, *i.e.*, documents that contain several modalities (*e.g.*, text, images, sound) requires taking into account all modalities, since they contain complementary pieces of information. The problem is that the various modalities are only weakly synchronized, they do not have the same rate and combining the information that can be extracted from them is not obvious. Of course, we would like to find generic ways to combine these pieces of information. Stochastic models appear as a well-dedicated tool for such combinations, especially for image and sound information.

Markov models are composed of a set of states, of transition probabilities between these states and of emission probabilities that provide the probability to emit a given symbol at a given state. Such models allow generating sequences. Starting from an initial state, they iteratively emit a symbol and then switch in a subsequent state according to the respective probability distributions. These models can be used in an indirect way. Given a sequence of symbols (called observations), hidden Markov models (HMMs, [87]) aim at finding the best sequence of states that can explain this sequence. The Viterbi algorithm provides an optimal solution to this problem.

For such HMMs, the structure and probability distributions need to be a priori determined. They can be fixed manually (this is the case for the structure: number of states and their topology), or estimated from example data (this is often the case for the probability distributions). Given a document, such an HMM can be used to retrieve its structure from the features that can be extracted. As a matter of fact, these models allow an audiovisual analysis of the videos, the symbols being composed of a video and an audio component.

Two of the main drawbacks of the HMMs is that they can only emit a unique symbol per state, and that they imply that the duration in a given state follows an exponential distribution. Such drawbacks can be circumvented by segment models [86]. These models are an extension of HMMs were each state can emit several symbols and contains a duration model that governs the number of symbols emitted (or observed) for this state. Such a scheme allows us to process features at different rates.

Bayesian networks are an even more general model family. Static Bayesian networks [80] are composed of a set of random variables linked by edges indicating their conditional dependency. Such models allow us to learn from example data the distributions and links between the variables. A key point is that both the network structure and the distributions of the variables can be learned. As such, these networks are difficult to use in the case of temporal phenomena. Dynamic Bayesian [85] networks are a generalization of the previous models. Such networks are composed of an elementary network that is replicated at each time stamp. Duration variable can be added in order to provide some flexibility on the time processing, like it was the case with segment models. While HMMs and segment models are well suited for dense segmentation of video streams, Bayesian networks offer better capabilities for sparse event detection. Defining a trash state that corresponds to non event segments is a well known problem is speech recognition: computing the observation probabilities in such a state is very difficult.

3.4. Multidimensional indexing techniques

Techniques for indexing multimedia data are needed to preserve the efficiency of search processes as soon as the data to search in becomes large in volume and/or in dimension. These techniques aim at reducing the number of I/Os and CPU cycles needed to perform a search. Multi-dimensional indexing methods either perform exact nearest neighbor (NN) searches or approximate NN-search schemes. Often, approximate techniques are faster as speed is traded off against accuracy.

Traditional multidimensional indexing techniques typically group high dimensional features vectors into cells. At querying time, few such cells are selected for searching, which, in turn, provides performance as each cell contains a limited number of vectors [79]. Cell construction strategies can be classified in two broad categories: *data-partitioning* indexing methods that divide the data space according to the distribution of data, and *space-partitioning* indexing methods that divide the data space along predefined lines and store each descriptor in the appropriate cell.

Unfortunately, the “curse of dimensionality” problem strongly impacts the performance of many techniques. Some approaches address this problem by simply relying on dimensionality reduction techniques. Other approaches abort the search process early, after having accessed an arbitrary and predetermined number of cells. Some other approaches improve their performance by considering approximations of cells (with respect to their true geometry for example).

Recently, several approaches make use of quantization operations. This, somehow, transforms costly nearest neighbor searches in multidimensional space into efficient uni-dimensional accesses. One seminal approach, the LSH technique [81], uses a structured scalar quantizer made of projections on segmented random lines, acting as spatial locality sensitive hash-functions. In this approach, several hash functions are used such that co-located vectors are likely to collide in buckets. Other approaches use unstructured quantization schemes, sometimes together with a vector aggregation mechanism [89] to boost performance.

3.5. Data mining methods

Data Mining (DM) is the core of knowledge discovery in databases whatever the contents of the databases are. Here, we focus on some aspects of DM we use to describe documents and to retrieve information. There are two major goals to DM: description and prediction. The descriptive part includes unsupervised and visualization aspects while prediction is often referred to as supervised mining.

The description step very often includes feature extraction and dimensional reduction. As we deal mainly with contingency tables crossing "documents and words", we intensively use factorial correspondence analysis. "Documents" in this context can be a text as well as an image.

Correspondence analysis is a descriptive/exploratory technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns. The results provide information, which is similar in nature to those produced by factor analysis techniques, and they allow one to explore the structure of categorical variables included in the table. The most common kind of table of this type is the two-way frequency cross-tabulation table. There are several parallels in interpretation between correspondence analysis and factor analysis: suppose one could find a lower-dimensional space, in which to position the row points in a manner that retains all, or almost all, of the information about the differences between the rows. One could then present all information about the similarities between the rows in a simple 1, 2, or 3-dimensional graph. The presentation and interpretation of very large tables could greatly benefit from the simplification that can be achieved via correspondence analysis (CA).

One of the most important concepts in CA is inertia, *i.e.*, the dispersion of either row points or column points around their gravity center. The inertia is linked to the total Pearson χ^2 for the two-way table. Some rows and/or some columns will be more important due to their quality in a reduced dimensional space and their relative inertia. The quality of a point represents the proportion of the contribution of that point to the overall inertia that can be accounted for by the chosen number of dimensions. However, it does not indicate whether or not, and to what extent, the respective point does in fact contribute to the overall inertia (χ^2 value). The relative inertia represents the proportion of the total inertia accounted for by the respective point, and it is independent of the number of dimensions chosen by the user. We use the relative inertia and quality of points to characterize clusters of documents. The outputs of CA are generally very large. At this step, we use different visualization methods to focus on the most important results of the analysis.

In the supervised classification task, a lot of algorithms can be used; the most popular ones are the decision trees and more recently the Support Vector Machines (SVM). SVMs provide very good results in supervised classification but they are used as "black boxes" (their results are difficult to explain). We use graphical methods to help the user understanding the SVM results, based on the data distribution according to the distance to the separating boundary computed by the SVM and another visualization method (like scatter matrices or parallel coordinates) to try to explain this boundary. Other drawbacks of SVM algorithms are their computational cost and large memory requirement to deal with very large datasets. We have developed a set of incremental and parallel SVM algorithms to classify very large datasets on standard computers.

WILLOW Project-Team

3. Scientific Foundations

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 [1] for the corresponding software (PMVS, http://grail.cs.washington.edu/software/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.

for free for academics, and licensing negotiations with several companies are under way.

Our recent work (Russel *et al.*, 2011) has 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. This direction is currently being continued in the PhD work of Mathieu Aubry. Our other current work outlined in detail in Section 6.1 is focused on (i) recovering indoor scene geometry from observations of person-object interactions video, (ii) visual place recognition in structured databases, where images are geotagged and organized in a graph, and (iii) developing a discriminative clustering approach able to discover geographically representative image elements from Google Street View imagery using only weak geographic supervision.

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, outlined in detail in Section 6.2), focuses on the two problems described next.

[1] The patent "Match, Expand, and Filter Technique for Multi-View Stereopsis," was issued December 11, 2012 and assigned patent number 8,331,615.

3.2.1. Learning image and object models.

Learning sparse representations of images has been the topic of much recent research. It has been used for instance for image restoration (e.g., Mairal *et al.*, 2007) and it has been generalized to discriminative image understanding tasks such as texture segmentation, category-level edge selection and image classification (Mairal *et al.*, 2008). We have also developed fast and scalable optimization methods for learning the sparse image representations, and developed a software called SPAMS (SPArse Modelling Software) presented in Section 5.2 . The work of J. Mairal is summarized in his thesis (Mairal, 2010). The most recent work has focused on developing a general formulation for supervised dictionary learning and investigating methods to learn better mid-level features for recognition.

3.2.2. Category-level object/scene recognition and segmentation

Another significant strand of our research has focused on the extremely challenging goals of category-level object/scene recognition and segmentation. Towards these goals, we have developed: (i) strongly-supervised deformable part-based model for object recognition and localization, (ii) a MRF model for segmentation of text in natural scenes, and (iii) algorithms for multi-class cosegmentation using a novel energy-minimization approach based on the developed convex relaxation for weakly supervised classifiers.

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.

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, outlined in detail in Section 6.3 , has focused on (i) developing a geometrical model for removing image blur due to camera shake, (ii) preparing an online image deblurring demo, and (iii) developing new formulation for image deblurring cast as a multi-label energy minimization problem.

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 6.4 .

3.4.1. 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. To this end we recently explored automatic mining of scene and action categories. Within PhD of Piotr Bojanowski we are currently extending this work towards exploiting richer textual descriptions of human actions and using them for learning more powerful contextual models of human actions in video.

3.4.2. 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. We explore the ways of enriching standard bag-of-feature representations with the higher-level information on objects, scenes and primitive human actions pre-learned on related tasks. We also investigate highly-efficient methods for computing video features motivated by the need of processing very large and increasing amounts of video.

3.4.3. Crowd characterization in video

Human crowds are characterized by distinct visual appearance and require appropriate tools for their analysis. In our work we develop generic methods for crowd analysis in video aiming to address multiple tasks such as (i) crowd density estimation and localization, (ii) characterization and recognition of crowd behaviours (e.g a person running against the crowd flow) as well as (iii) detection and tracking of individual people in the crowd. We address the challenge of analyzing crowds under the large variation in crowd density, video resolution and scene structure.

3.4.4. Modeling and recognizing person-object and person-scene interactions.

Actions of people are tightly coupled with their environments and surrounding objects. Moreover, object function can be learned and recognized from observations of person-object interactions in video and still images. Designing and learning models for person-object interactions, however, is a challenging task due to both (i) the huge variability in visual appearance and (ii) the lack of corresponding annotations. We address this problem by developing weakly-supervised techniques enabling learning interaction models from long-term observations of people in natural indoor video scenes such as obtained from time-lapse videos on YouTube. We also explore stereoscopic information in 3D movies to learn better models for people in video including person detection, segmentation, pose estimation, tracking and action recognition.