



Project acronym	SIMBAD
Project full title	Beyond Features: Similarity-Based Pattern Analysis and Recognition
Deliverable Responsible	Dipartimento di Informatica Università Ca' Foscari di Venezia Via Torino 155 30172 Venezia Mestre Italy http://www.dsi.unive.it/~pelillo
Project web site	http://simbad-fp7.eu
EC project officer	Teresa De Martino
Document title	First Periodic Report
Deliverable	D1.2
Document type	Report
Dissemination level	Public
Contractual date of delivery	M 13
Project reference number	213250
Status & version	Definitive
Work package	WP 1
Deliverable responsible	UNIVE
Contributing Partners	UNYORK, TUD, IST, UNIVR, ETH Zurich
Author(s)	Marcello Pelillo
Additional contributor(s)	Veronica Giove

PROJECT PERIODIC REPORT

Grant Agreement number: 213250

Project acronym: SIMBAD

Project title: Beyond Features: Similarity-Based Pattern Analysis and Recognition

Funding Scheme: Collaborative Project

Date of latest version of Annex I against which the assessment will be made: 01/02/2008

Periodic report: 1st 2nd 3rd 4th

Period covered: from 01/04/2008 to 31/03/2009

**Name, title and organisation of the scientific representative of the project's coordinator:
Marcello Pelillo, Associate Professor, Computer Science Department, Università Ca' Foscari
di Venezia**

Tel: +39 041 234 8440

Fax: +39 041 234 8419

E-mail: E-mail: pelillo@dsi.unive.it

Project website address: <http://simbad-fp7.eu>

Declaration by the scientific representative of the project coordinator

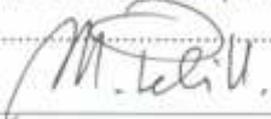
I, as scientific representative of the coordinator of this project and in line with the obligations as stated in Article II.2.3 of the Grant Agreement declare that:

- The attached periodic report represents an accurate description of the work carried out in this project for this reporting period;
- The project (tick as appropriate):
 - has fully achieved its objectives and technical goals for the period;
 - has achieved most of its objectives and technical goals for the period with relatively minor deviations;
 - has failed to achieve critical objectives and/or is not at all on schedule.
- The public website is up to date.
- To my best knowledge, the financial statements which are being submitted as part of this report are in line with the actual work carried out and are consistent with the report on the resources used for the project (section 6).
- All beneficiaries, in particular non-profit public bodies, secondary and higher education establishments, research organisations and SMEs, have declared to have verified their legal status. Any changes have been reported under section 5 (Project Management) in accordance with Article II.3.f of the Grant Agreement.

Name of scientific representative of the Coordinator: Marcello Pelillo.....

Date: 5 / 6 / 2009

Signature of scientific representative of the Coordinator:



1. Publishable summary

The field of pattern recognition (or machine learning) is concerned with the automatic discovery of regularities in data through the use of computer algorithms, and with the use of these regularities to take actions such as classifying data into different categories, with a view to endow artificial systems with the ability to improve their own performance in the light of new external stimuli. This ability is instrumental in building next-generation artificial cognitive systems (ACS's), namely systems that can perceive, reason and interact robustly in open-ended environments. The socio-economic implications of this scientific endeavour are enormous, as ACS's will have applications in a wide variety of real-world scenarios ranging from industrial manufacturing to vehicle control and traffic safety, to remote and on-site (environmental) sensing and monitoring, and to medical diagnostics and therapeutics.

This project aims at bringing to full maturation a paradigm shift that is currently just emerging within the pattern recognition and machine learning domains, where researchers are becoming increasingly aware of the importance of similarity information *per se*, as opposed to the classical feature-based (or vectorial) approach. Indeed, the notion of similarity has long been recognized to lie at the very heart of human cognitive processes and can be considered as a connection between perception and higher-level knowledge, a crucial factor in the process of human recognition and categorization.

Traditional pattern recognition techniques are centered around the notion of "feature". According to this view, each object is described in terms of a vector of numerical attributes and is therefore mapped to a point in a Euclidean (geometric) vector space so that the distances between the points reflect the observed (dis)similarities between the respective objects. This kind of representation is attractive because geometric spaces offer powerful analytical as well as computational tools that are simply not available in other representations. Indeed, classical pattern recognition methods are tightly related to geometrical concepts and numerous powerful tools have been developed during the last few decades, starting from linear discriminant analysis in the 1920's, to perceptrons in the 1960's, to kernel machines in the 1990's.

The geometric approach suffers from a major intrinsic limitation, which concerns the representational power of vectorial, feature-based descriptions. In fact, there are numerous application domains where either it is not possible to find satisfactory features or they are inefficient for learning purposes. This is typically the case when experts cannot define features in a straightforward way, when data are high dimensional, when features consist of both numerical and categorical variables, and in the presence of missing or inhomogeneous data. But, probably, this situation arises most commonly when objects are described in terms of structural properties, such as parts and relations between parts, as is the case in shape recognition. This led in 1960's to the development of the structural pattern recognition approach, which uses symbolic data structures, such as strings, trees, and graphs for the representation of individual patterns, thereby, reformulating the recognition problem as a pattern-matching problem.

In the last few years, interest around purely similarity-based techniques has grown considerably. For example, within the supervised learning paradigm (where expert-labeled training data is assumed to be available) the now famous "kernel trick" shifts the focus from the choice of an appropriate set of features to

the choice of a suitable kernel, which is related to object similarities. However, this shift of focus is only partial, as the classical interpretation of the notion of a kernel is that it provides an implicit transformation of the feature space rather than a purely similarity-based representation. Similarly, in the unsupervised domain, there has been an increasing interest around pairwise algorithms, such as spectral and graph-theoretic clustering methods, which avoid the use of features altogether.

Despite its potential, presently the similarity-based approach is far from seriously challenging the traditional paradigm. This is due mainly to the sporadicity and heterogeneity of the techniques proposed so far and the lack of a unifying perspective. On the other hand, classical approaches are inherently unable to deal satisfactorily with the complexity and richness arising in many real-world situations. This state of affairs hinders the application of machine learning techniques to a whole variety of relevant, real-world problems. Hence, progress in similarity-based approaches will surely be beneficial for machine learning as a whole and, consequently, for the long-term enterprise of building ACS's. However, by departing from vector-space representations one is confronted with the challenging problem of dealing with (dis)similarities that do not necessarily possess the Euclidean behavior or not even obey the requirements of a metric. The lack of the Euclidean and/or metric properties undermines the very foundations of traditional pattern recognition theories and algorithms, and poses totally new theoretical/computational questions and challenges that we are addressing with this project.

With this project, we are undertaking a thorough study of several aspects of similarity-based pattern analysis and recognition methods, from the theoretical, computational, and applicative perspective, with a view to substantially advance the state of the art in the field, and contribute towards the long-term goal of organizing this emerging field into a more coherent whole. The whole project revolves around two main themes, which basically correspond to the two fundamental questions that arise when abandoning the realm of vectorial, feature-based representations, namely:

- How can one *obtain* suitable similarity information from object representations that are more powerful than, or simply different from, the vectorial?
- How can one *use* similarity information in order to perform learning and classification tasks?

Although the two issues are clearly interrelated, it is advantageous to keep them apart as this allows one to separate the *similarity generation* process (a data modeling issue) from the *learning and classification* processes (a task modeling issue). According to this perspective, the very notion of similarity becomes the pivot of non-vectorial pattern recognition in much the same way as the notion of feature-vector plays the role of the pivot in the classical (geometric) paradigm. This results in a useful modularity, which means that all interactions between the object representation and the learning algorithm are mediated by the similarities, which is where the domain knowledge comes into the scene.

During the first year of the project we started working on the above fundamental questions. As for the first question, we aimed at devising suitable similarity measures for non-vectorial data, specifically tailored to a given task. We have focussed primarily on structured data (e.g., graphs), because of their expressive power and ubiquity, and on geometric measures as they allow one to employ the whole arsenal of powerful techniques available in the (classical) pattern recognition literature. We have also explored an alternative to this "tailoring" approach, which consists in learning similarities directly from training data.

As concerns the second question, we have both addressed foundational issues related to similarity information and developed practical similarity-based algorithms that do not depend on the actual object representation. In particular, as concerns the latter objective, we have distinguished between the situation where the informational content associated with the violation of the geometric properties is limited, or is simply an artifact of the measurement process, and that where this is not the case. This distinction is important as, depending on the actual situation, two complementary strategies will be pursued: the first attempts to impose geometricity by somehow transforming or re-interpreting the similarity data, the second does not and works directly on the original similarities. Our theoretical analysis is expected to provide means to distinguish between these cases, thereby allowing one to understand which of the two approaches is more appropriate for the problem at hand.

Apart from the potential applications mentioned before, within SIMBAD we shall devote in the next two years substantial effort towards tackling two large-scale biomedical imaging applications. With the direct involvement of leading pathologists and neuroscientists from the University Hospital Zurich and the Verona-Udine Brain Imaging and Neuropsychology Program, we expect to contribute towards the concrete objective of providing effective, advanced techniques to assist in the diagnosis of renal cell carcinoma, one of the ten most frequent malignancies in Western countries, as well as of major psychoses such as schizophrenia and bipolar disorder. These problems are not amenable to be tackled with traditional machine learning techniques due to the difficulty of deriving suitable feature-based descriptions. Indeed, many biomedical applications exhibit precisely the same characteristics. Hence, a successful outcome of our experimentation would provide evidence as to the practical applicability of our approach in biomedicine, thereby fostering further research along the lines set up by SIMBAD, both at the methodological and at the practical level. This would potentially open new opportunities in health and disease management and bring radical improvements to the quality and efficiency of our healthcare systems. Although in our original plans this activity was supposed to start at the second year of the project, we have anticipated it by a few months.

SIMBAD involves the following partners:

- UNIVE: Ca' Foscari University (Venice, Italy), *coordinator*
- UNİYORK: University of York (England)
- TUD: Delft University of Technology (The Netherlands)
- IST: Instituto Superior Técnico (Lisbon, Portugal)
- UNIVR: University of Verona (Italy)
- ETH Zurich: Eidgenössische Technische Hochschule Zürich (Switzerland)

The consortium has been carefully designed so as to include top-level competences in all relevant areas and problems in pattern recognition. In addition, we strove to set up a highly cohesive network where each research unit not only is well characterized in terms of competences, problems addressed, methodologies used and objectives, but is also tightly coupled with the rest of the network. All members in the consortium have an established international reputation and a long experience in the fields of machine learning and pattern recognition, where in the past few years have also contributed substantially to advance the state-of-the-art of the similarity-based paradigm.

The complementarity of expertises is crucial for our endeavour, as it allows us to attack each problem from different standpoints, thereby fostering cross-fertilization of ideas.

The SIMBAD website is: <http://simbad-fp7.eu>

2. Project objectives for the period

In this project we aim at advancing the state of the art in similarity-based pattern analysis and recognition from the theoretical, computational, and applicative perspective, and contributing towards the long-term goal of organizing this emerging field into a more coherent whole. As outlined in the “Publishable summary”, the project revolves around two main themes, which concerns the issues of how to obtain suitable similarity information from non-vectorial representations, and how to use them, irrespective of the way in which they are obtained. In addition to these two basic themes, a third one arises which concerns the validation of the proposed techniques and their applicability to real-world problems. Pattern recognition is intrinsically an application-based field with well-established validation methodologies. These will be used to quantitatively evaluate the success of the proposed research on large-scale applications with clear social impact.

Accordingly, the objectives for the first year of the project have been structured around the following three strands.

1. Deriving similarities for non-vectorial data. The goal here is to develop suitable similarity measures for non-vectorial data. We have focussed primarily on structured data (e.g., graphs), because of their expressive power and ubiquity, and on geometric measures as they allow one to employ the whole arsenal of powerful techniques available in the geometric pattern recognition literature. We have pursued our goal by investigating the use of generative models to capture the connectivity relations encoded in the graphs and to infer the topological relations between points in a non-Euclidean manifold. In this context, we have also analyzed kernels that do not satisfy the Mercer condition in an attempt to provide statistical characterizations of its effectiveness. With a view to devising more descriptive structural kernels, we have also investigated dissimilarity functions based on information theory. An alternative to this “kernel tailoring” approach consists in learning good similarities directly from training data. In the first year, we have explored mainly ideas based on ensemble-clustering. In the second year will shall explore ideas from game theory.

2. Learning and classification with non-(geo)metric similarities. Within this research strand we aim at both addressing foundational issues related to similarity information and developing practical algorithms that do not depend on the actual object representation. In particular, as concerns the latter objective, we have distinguished two cases, which in turn lead to two complementary approaches. On the one hand, we have considered the case where the informational content of non-(geo)metricity is limited or caused by measurement error. In this case it is a plausible strategy to perform some correction on the similarity data (or, alternatively, finding an alternative vectorial representation) in an attempt to impose (geo)metricity, and then use classical geometric techniques. On the other hand, when the information content of non-(geo)metricity is relevant one needs brand new tools, as standard techniques would not work in this case.

More specifically, the activity within these themes has been organized around three main areas:

a) *Foundations of non (geo)metric similarities.* One of the first objectives here was to explore the causes and originations of non-(geo)metric data. In particular, we aimed to investigate how these causes influence the computability of various classifiers and their performances. To our knowledge, systematic investigations of the reasons and circumstances under which deviations from the standard Euclidean metric arise have not been yet undertaken. Consequently, it is also unknown how the various options of handling the data affect the result in relation with the cause of its non-metric nature. Besides investigating the origins of non-(geo)metricity, we studied the actual information content of such violations, i.e., their influence on standard machine-learning algorithms that were originally designed for Euclidean data. Based on previous results showing that certain clustering procedures are essentially “blind” against violations of the triangle inequality, we aimed at finding similar robustness results for more advanced classification methods.

b) *Imposing geometricity on non-geometric similarities (embedding).* Despite the growing interest around embedding, the search for robust embeddings procedures on structured data such as weighted graphs has proven elusive, and their geometric and probabilistic characterizations still remains to be explored in depth. To this end, we used ideas from spectral geometry to extract differential invariants from the graph-spectra. With the geometric characterization to hand, we aim to construct probabilistic models that can account for the distributions of the invariants. Within this strand, we shall also investigate approaches that, instead of approximating the original (dis)similarities by Euclidean distances, try to preserve the underlying group structure of the data, thereby bringing us back to the geometric domain and hence allowing us to apply standard methods.

c) *Learning with non-(geo)metric similarities.* When there is significant information content in the non-(geo)metricity of the data, it is hard to define a single, global objective function that satisfactorily accounts for the complexity of the problem at hand and hence alternative approaches are needed. Game theory was developed precisely to overcome the limitations of single-objective optimization as it aims at modeling complex situations where players make decisions in an attempt to maximize their own (mutually conflicting) returns. Its main point is to shift the emphasis from (global) optimality criteria to equilibrium conditions. In the last six months of the project we started a systematic study with a view to *modeling* generic pattern recognition problems in terms of purely game-theoretic concepts, based on the idea of game-theoretic formalization of the competition between the hypotheses of class membership.

3. Validation. To our best knowledge, there does not exist any well-established testbed for assessing the quality of similarity-based learning approaches. Given the heterogeneity of different approaches that we are pursuing in this project, it is of particular importance to build a real-world testbed that specifically addresses the various difficulties involved with non-metric data. At the same time, however, this testbed must be diverse enough to ensure generalization of learning methods and to avoid over-fitting of single databases. To this end, we will focus on biomedical datasets that nicely combine high practical relevance of the underlying learning tasks and intrinsically non-metric dissimilarity data. Although our original plan was to start the activities related to these applications in the second year, we have somewhat anticipated this work (see below).

Following the outline set forth above, the three major themes form the basis for structuring the project's work plan into coherent work packages (WP's). Specifically, WP2 covers the topics of deriving similarities for non-vectorial data (theme 1), while the second theme concerning learning and classification with non-(geo)metric similarities is addressed by WP3, WP4 and WP5. Finally, the validation phase is undertaken in WP6 and WP7. An additional work package (WP1) deals with management issues, while WP8 deals with dissemination strategies.

More specifically, as far as the scientific and the dissemination part is concerned, the project is articulated in the following work packages:

WP2. Deriving similarities for non-vectorial data (structural kernels)

WP3. Foundations of non-(geo)metric similarities

WP4. Imposing geometricity on non-geometric similarities (embedding)

WP5. Learning with non-(geo)metric similarities

WP6. Analysis of tissue micro-array (TMA) images of renal cell carcinoma

WP7. Analysis of brain magnetic resonance (MR) scans for the diagnosis of mental illness

WP8: Dissemination, communication and exploitation

In the next section we shall describe the activities done within each work package (for more details we refer to related publications and/or deliverables).

3. *Work progress and achievements during the period (months 1–12)*

Work package WP2:

Deriving similarities for non-vectorial data (structural kernels)

Work package leader: IST

Participants: IST, UNIVR, UNIVE, UNİYORK, ETH Zurich

Start month: 1

End month: 24

Overall person-months: 52

The main objective of the work package WP2 (“Deriving similarities for non-vectorial data”) is to develop computational models for generating kernels and more general similarity measures for non-vectorial data. It is divided into three main tasks: WP2.1 (“Generative Kernels”), WP2.2 (“Compression kernels”), and WP2.3 (“Learning and Combining Similarities”). In this report, we briefly describe the main achievements in the context of each of these tasks, and point to the relevant publications where these achievements are described in greater detail.

Generative Kernels

Generative kernels and information-theoretic kernels (the topic of WP2.2) are closely related. Both are based on the assumption that the objects of interest were generated by a probabilistic mechanism - a source, in information theoretic terms - and then proceed by defining (dis)similarity measures or kernels between models of these probabilistic sources. In fact, information-theoretic kernels can be considered as generative kernels, although the information-theoretic perspective tends to be more agnostic; i.e., it does not assume that the adopted model reflects the truth. In this section, we will describe the work on the topic traditionally called “generative kernels”.

To serve as background material, we started by producing a technical report with all the necessary mathematical foundations, as well as a review of generative and information-theoretic kernels. Apart from this technical report, which was produced by the IST group, the main contributor to the WP2.1 task was the UNIVR partner. The research work has mainly focused on the following two sub-topics: *Fisher kernels* and *score spaces*.

The Fisher kernel is the earliest and best known generative kernel. The UNIVR group has therefore conducted a preliminary study, including several experiments with the goal of assessing the suitability of Fisher kernels to several applications, such as 2D shape recognition, gesture recognition, speaker classification, and EEG signal classification. Given the sequential nature of this type of data, hidden Markov models (HMM) have been a natural choice as generative models. Moreover, a study of the best procedure to build generative models for Fisher kernels was carried out. In particular, we proposed a new way to fit

generative models to the training data from each class. The idea is to first perform model clustering on the unlabeled data in order to discover the optimal structure for the entire sample set. Then, the label information is retrieved and generative scores are computed. Comparative tests provide a preliminary assessment of the merits of this approach and suggest further developments. This work was described in detail in the publication [Hidden 1].

The UNIVR group has also explored the topic of score spaces, also known as generative embeddings. Score spaces are highly informative spaces where we can project sample objects when using generative models to extract fixed-size feature vectors (the so-called scores). In such spaces, the inner product may correspond to a kernel, thus kernel-based discriminative classifiers may be used. In this context, during the first year of SIMBAD, the UNIVR group produced two contributions, both employing HMMs, and which are described in the next two paragraphs.

Firstly, we proposed to derive the features characterizing each sequence from specific components of the individual HMMs. Every feature measures the relevance of a specific component (like states or transitions) of a given HMM when an input sequence is fed to this HMM, or, better, how a specific component contributes to the explanation of the sequence. A thorough experimental evaluation showed that this class of methods is robust and may drastically improve over the standard HMM-based schemes. This is particularly evident when the original models are insufficient to solve the problem, usually due to small training sets, incorrect model topology, or bad modeling assumptions. This line of work has been described in detail in publication [Bicego, Pekalska, Tax, and Duin, 2009].

Secondly, we worked on a novel score space that captures the generative process by encoding it in an entropic feature vector. In this way, both uncertainty in the generative model learning step and “local” agreement of data observations with respect to the generative process can be represented. The idea is to capture the difference in the generative process between the samples; to do this we pool together the samples and, once the generative model is learnt, we calculate a distance between the statistics collected over all the samples, encoded in the parameter estimate, and the individual j th sample statistics $p(y|x(j))$, where y represent the set of hidden variables, possibly containing the class variable. These methods showed promising results, outperforming the Fisher scores over the chicken pieces dataset. This results are described in detail in [Perina, Cristani, Castellani, and Murino, 2009].

Following the direction described in the previous paragraph, we are currently working on another score space that exploits the sum factorization of the free energy associated to a generative model. Free energy is a popular score function, which is minimized in variational model training, representing a lower bound on the negative log-likelihood of the visible variables. Making a discriminative use of these “pieces” of free energy, we can see which parts of a generative model are more useful for discrimination, theoretically outperforming standard generative classification which trivially can be viewed as a special case of our method. Preliminary results over a variety of generative models (hidden Markov models, mixture of Gaussians, probabilistic index maps, latent Dirichlet allocation) testify for the feasibility of the approach, as reported in [Hidden 2].

Compression kernels

This task, which would be better called “Information theoretic kernels”, aims at devising ways to obtain kernels for non-vectorial data, based on information theory. More specifically, the idea is to assume that the objects (to be clustered or classified) were generated by some probabilistic source, and then define kernels between source models. More formally, by defining a family S containing the distributions from which the data points (in the input space X) are assumed to have been generated, and defining a map from X from S (e.g., via maximum likelihood estimation), a distribution in S may be fitted to each datum. Therefore, a kernel defined on $S \times S$ induces a kernel on $X \times X$, via map composition. In text categorization, this approach is an alternative to the Euclidean geometry inherent to the bag-of-words representations. In fact, approaches that map data to statistical manifolds, equipped with well motivated non-Euclidean metrics, often outperform SVM classifiers with linear kernels. Some of these kernels have a natural information theoretic interpretation, thus bridging between kernel methods and information theory.

The idea in “compression kernels” is to bypass the explicit construction of the map from X from S . In principle, it is possible to obtain the entropy of a source by finding the best compression rate achievable for it, thus bypassing the estimation of an explicit model source. Of course this requires the use of a universal source coding technique (e.g., of the Lempel-Ziv class), and the result is asymptotic. The same idea can be applied to non-parametrically obtain dissimilarities (divergences) between sources, as was proposed by Ziv and Merhav (1993). Before fully embarking on the task of obtaining compression based kernels (although we did obtain some progress in that direction, as reported below), we have decided to complete our work on information-theoretic kernels, as explained in the next paragraphs.

The main goal of this line of work was to obtain a wider family of information-theoretic kernels, containing the previously known ones as particular cases. We did that by introducing a class of kernels using nonextensive information theory, which contains previous information theoretic kernels as particular elements. The famous Shannon and Rényi entropies share the so-called extensivity property: the joint entropy of two independent random variables equals the sum of their entropies. Dropping this property yields the so-called nonextensive entropies (Havrda and Charvát, 1967; Lindhard and Nielsen, 1971; Tsallis, 1988), which have raised interest among physicists for modeling phenomena with long-range interactions, and in constructing nonextensive generalizations of the Boltzmann-Gibbs statistical mechanics (Abe, 2006). Nonextensive entropies have been recently used in signal/image processing (Li, Fan, and Li 2006) and other areas (Gell-Mann and Tsallis, 2004). The so-called Tsallis entropies (Havrda and Charvát, 1967; Tsallis, 1988) form a parametric family of nonextensive entropies that includes the Shannon entropy as a particular case.

Our main achievements were the following:

- The concept of q -convexity, generalizing that of convexity, for which we prove a Jensen q -inequality. The related concept of Jensen q -differences, which generalize Jensen differences, is also proposed. Based on these concepts, we introduce the Jensen-Tsallis (JT) q -difference, a nonextensive generalization of the JS divergence, which is also a “mutual information” in the sense of Furuichi (2006).
- Characterization of the JT q -difference, with respect to convexity and extrema, extending work by Burbea and Rao (1982) and by Lin (1991) for the JS divergence.

- Definition of k-th order joint and conditional JT q-differences for families of stochastic processes, and derivation of a chain rule.
- A broad family of (nonextensive information theoretic) positive definite kernels, interpretable as nonextensive mutual information kernels, ranging from the Boolean to the linear kernels, and including the JS kernel proposed by Hein and Bousquet (2005).
- A family of (nonextensive information theoretic) positive definite kernels between stochastic processes, subsuming well-known string kernels (e.g., the p-spectrum kernel) (Leslie, Eskin, and Noble, 2002).
- Extensions of results of Hein and Bousquet (2005) proving positive definiteness of kernels based on the unbalanced JS divergence. A connection between these new kernels and those studied by Fuglede (2005) and Hein and Bousquet (2005) is also established. In passing, we show that the parametrix approximation of the multinomial diffusion kernel (Lafferty and Lebanon, 2005) is not positive definite.

All these results on nonextensive information-theoretic kernels were described in full detail in the following publications (found in the annex): [Martins, Aguiar, Figueiredo, 2008], [Martins, Figueiredo, Aguiar, Smith, Xing, 2008], and [Martins, Smith, Xing, Aguiar, Figueiredo, 2009].

Learning and Combining similarities

This task is devoted to the problem of deriving similarities from examples and combining them. In the first year of the project we explored ideas from clustering ensembles.

Learning Pairwise Similarity

Each clustering algorithm induces a similarity between data points, according to the underlying clustering criterion. Fred and Jain (2006) proposed a cluster ensemble (CE) approach for learning pairwise similarities between objects. In that work, the quality of clusters is assessed using a stability measure, used for the selection of meaningful clusters for which the underlying clustering criterion may contribute to the overall estimation of the pairwise similarities. The use of different parameters enabled the derivation of similarities between patterns without a priori information about the number of clusters or tuning of parameter values. Furthermore, using average link clustering with a lifetime criterion over the learned pairwise similarities on several artificial and real datasets showed the robustness and usefulness of the approach in revealing the underlying cluster structure in the data.

In the context of SIMBAD, the CE framework was further explored and extended to learn similarity relations of temporal data, with application to the identification of stress states from ECG signals. Taking as motivating application the evaluation of changes in ECG morphology in the course of a stress-inducing computer-based activity, the *evidence accumulation clustering* (EAC) method was applied and evaluated using different clustering algorithms for the construction of clustering ensembles as well as various

algorithms for final extraction of the (combined) final partition; these various setups were additionally explored in conjunction with feature selection and feature extraction techniques.

The developed work presents several innovative aspects:

- *Stress-related ECG morphological changes.* In previous work, stress has been found to be associated with heart rate variability. However, morphological changes have not been studied so far. In our work, we addressed this issue, by assessing the temporal evolution of ECG morphology, summarized in a similarity matrix between heart beat waves, indices of the matrix corresponding to increasing time stamps. Our results confirm this morphology change hypothesis, showing clear dissimilarity between ECG patterns at the beginning and at the end of the task; furthermore, by clustering the learned similarity matrix using the CE approach, such a hypothesis is confirmed by revealing distinct clusters.
- *Methodology for the analysis of temporal data based on the clustering ensemble approach,* which, to our knowledge is also new under the unsupervised learning paradigm.
- *Genetic algorithm for temporal data denoising.* Clustering of stationary temporal data with abrupt changes in the temporal organization model is a relatively simple problem. Given the continuous time evolution of stress levels, clustering algorithms are deemed to fail to detect well separated groups of patterns. Therefore, elimination of samples that correspond to the continuous transition between distinct states (denoted as noisy patterns) is one possible approach to detect if such meaningful distinct clusters are present in the data. The genetic algorithm proposed identifies and eliminates transition time frames based on a cluster separability fitness function.

This work, supervised by Prof. Ana Fred, resulted in the MSc Thesis "Identification of Stress States from ECG Signals using Unsupervised Learning" [Medina, 2009], by Liliana Medina, summarized in the extended abstract attached [Medina and Fred, 2009]. The thesis has contributions to unsupervised learning of similarities, both from methodological and algorithmic points of view, that are foreseen to be soon further validated and published in international conferences and journals. Furthermore, the open issues raised by this work are foreseen to form a basis for international collaborations within the SIMBAD partners, in particular in assessing similarities between ECG signals and other electrophysiological data, without explicit feature extraction.

Constrained Clustering

Recent work on clustering has focused on the incorporation of *a priori* knowledge, mostly in the form of pairwise constraints, aiming to improve clustering quality of individual clustering algorithms, and find appropriate clustering solutions to specific tasks or interests.

In the context of SIMBAD, we proposed to extend and integrate the constrained clustering idea into the CE framework. Such integration can be implemented at three main levels, and combinations thereof: (a) on the construction of the CE, by explicitly applying constrained clustering algorithms; (b) during information combination phase, by forcing (hard constraints) or encouraging (soft constraints) pairwise associations; (c) at the step of extraction of the final (combined) data partition. In the work developed so far, we proposed an

extension to EAC (termed CEAC), and a novel algorithm (ACCCS) to solve the CE problem using pairwise constraints (must-link and cannot-link). CEAC consists of enforcing the clustering algorithm, which produces the consensus partition from the learned similarity, to support the incorporation of must-link and cannot-link constraints. The ACCCS approach comprises the maximization of both the similarity between CE partitions and a target consensus partition, and constraints satisfaction. Experimental results using 4 synthetic and 8 real data sets have shown the proposed constrained clustering combination methods performances are superior to the unconstrained Evidence Accumulation Clustering. Part of this work is reported in [Duarte, Fred, and Duarte, 2009].

Dissimilarity Increments Statistics and Cluster Validity

Cluster validity is a key issue in unsupervised learning. In previous work [Fred, Leitão, 2003], a criterion was proposed with an underlying model-based characterization of inter-pattern relationships. According to this hypothesis, therein addressed and discussed, absolute differences of dissimilarity values (dissimilarity increments), computed between neighboring patterns within a natural cluster, follow an exponential distribution. In the context of SIMBAD, we built on this statistical model of dissimilarity increments, and addressed the problem of analyzing clustering solutions based on probabilistic attributed graphs. Assuming that the dissimilarity values computed between neighboring patterns within a natural cluster follow an exponential distribution, we proposed a generative model for the clusters. This formed the basis for the design of a new cluster validity index, consisting of the description length of the data partition, represented by a probabilistic attributed graph inferred from the data, conditioned on the given partition. Decision between clustering solutions based on the new index follows a *minimum description length* (MDL) criterion.

The proposed criterion was evaluated in three distinct scenarios: (a) selection of a design parameter for a given clustering algorithm; (b) selection between clustering solutions produced by different clustering algorithms; (c) the choice between combination results in a clustering ensemble approach. Results on several (synthetic and real) datasets, revealed good performance of the index in selecting a partition or design parameter. Part of this work was published in [Fred and Jain, 2009].

Ongoing work includes a further evaluation of the proposed criterion in a comparative study with cluster validity indices reported in the literature. Additionally, collaborative work with UNIVE is being undertaken in the harmonization and integration of the dissimilarity increments statistic with the concept of dominant sets, a graph-theoretic concept which generalizes that of a maximal clique to edge-weighted graphs, in order to derive new dissimilarity-based clustering strategies and algorithms.

Publications.

M. Bicego, E. Pekalska, D. Tax, R. Duin, "Component-based discriminative classification for hidden Markov models", *Pattern Recognition*, in press, 2009.

J. Duarte, A. Fred, F. Duarte, "Combining data clusterings with instance level constraints", *Proceedings of the 9th Intl. Workshop on Pattern Recognition in Information Systems*, Milan, Italy, May 2009.

A. Fred and A. K. Jain, "Cluster Validation using a Probabilistic Attributed Graph", *Proceedings of the International Conference on Pattern Recognition*, Tampa, FL, USA, December 2008.

A. Martins, P. Aguiar, and M. Figueiredo. "Tsallis kernels on measures" , *Proceedings of the IEEE Information Theory Workshop*, Porto, Portugal, 2008.

A. Martins, M. Figueiredo, P. Aguiar, N. Smith, and E. Xing. "Nonextensive entropic kernels", *Proceedings of the International Conference on Machine Learning – ICML'08*, Helsinki, inland, 2008.

A. Martins, N. Smith, E. Xing, P. Aguiar, and M. Figueiredo, "Nonextensive information-theoretic kernels on measures", *Journal of Machine Learning Research*, vol. 10, pp. 935-975, May 2009.

L. Medina and A. Fred, "Identification of Stress States from ECG Signals using Unsupervised Learning Methods", Extended Abstract of Master's Degree Dissertation on Electrical Engineering and Computers, Instituto Superior Técnico, April 2009.

L. Medina, "Identification of Stress States from ECG Signals using Unsupervised Learning Methods", Master's Degree Dissertation on Electrical Engineering and Computers, Instituto Superior Técnico, April 2009 (in Portuguese).

A. Perina, M. Cristani, U. Castellani, V. Murino, "A new generative feature set based on entropy distance for discriminative classification", *International Conference on Image Analysis and Processing – ICIAP'09*, Salerno, Italy, 2009 (accepted).

References [Hidden 1] and [Hidden 2] are not listed since they were submitted and are under double-blind review.

Work package WP3:

Foundations of non-(geo)metric similarities

Work package leader: TUD

Participants: TUD, ETH Zurich

Start month: 1 End month: 18 Overall person-months: 29

The objective of this workpackage is to study both the causes of the lack of (geo)metricity in the similarity data and its effects over traditional machine learning algorithms. The workpackage is divided into the following tasks: WP3.1 ("Causes and origination of non-(geo)metricity") and WP3.2 ("Information content of (geo)metricity violations").

Causes and origination of non-(geo)metricity

Our first goal in WP3.1 was to address fundamental questions pertaining to the intrinsic nature of non-(geo)metric data. Two major steps can be distinguished in the construction of recognition systems for pattern classes of real world objects. These are *representation* and *generalization*. The step of generalization has been well studied for the case of vector representations in Euclidean spaces. However, in the pattern recognition practice non-Euclidean dissimilarity measures and/or indefinite kernels (or similarity measures) are frequently used for representation. They implicitly describe objects in non-Euclidean vector spaces (determined through embeddings) for which generalization is less well defined.

There are three ways to handle this problem: (1) suitable adaptation of the (dis)similarity measure, (2) transformation of the non-Euclidean space into a Euclidean space via a correction procedure, and (3) extension of the set of generalization procedures to non-Euclidean spaces. Which solution is to be preferred may be related to the cause of the non-Euclidean relations between the objects in a particular problem.

In deliverable D3.1 (and in related publications) we have tried to analyze them on the basis of examples from the real world as well as artificial ones. Non-Euclidean behavior can arise either by non-intrinsic or intrinsic causes. The first ones are the result of the lack of either computational or observational power. The second ones are the consequence of an essential non-Euclidean judgment of the object dissimilarities, often resulting from restricted, pairwise comparisons. The report is concluded with a discussion on the possible identification of the cause of a non-Euclidean representation for the generalization step.

During the first year of the project various data sets have also been collected. They are not formally public yet as a way has to be defined how to make them public (format, annotations). In a further cooperation with Pekalska a set of Matlab tools will be created for analyzing these datasets and for use inside SIMBAD. This will be delivered later in the project.

Information content of (geo)metricity violations

Besides investigating the origins of non-metricity, in WP3.2 we have studied the actual information content of such violations, i.e., their influence on standard machine-learning algorithms originally designed for Euclidean data. Here, the theoretical studies have been concentrated on (i) a nonparametric Bayesian study of pairwise clustering and on (ii) an information theoretic approach to approximate optimization which will yield a fundamental basis for learning. Both topics are discussed in the deliverable by the ETH Zurich group.

The nonparametric Bayesian approach to pairwise clustering yields new insights into clustering models, which are invariant to constant shifts of the dissimilarities between objects. Such shifts transform non-metric dissimilarities in such a way that the cluster statistics is unchanged by the dissimilarity transformation. The Bayesian approach to graph based clustering models provides a generative view on these methods and allows us to identify and extend the invariance mechanisms to a larger class of cost functions than previously known.

The second study addresses the general problem of model validation which requires balancing the stability aspect of learning methods and the information content of models. The noise in the data source should have little influence on the final model and its parameters to be learned from data. On the other hand, too simplistic models show high stability at the expense of low information content. A comprehensive theory of model validation, therefore, has to find an optimal compromise between the two requirements. The ETH Zurich group has developed an information theoretic framework for model validation where approximate optimization by e.g. Gibbs sampling is used to code information. The codebook of code vectors in Shannon's theory of random coding is replaced by a set of code problems which are approximated up to an approximation precision of γ . Two instances of code problems, a training problem and a test problem are necessary to develop the code.

The framework is described in more detail in the deliverable D3.2. We consider this theory as a substantial contribution to the research theme of SIMBAD since the theory differentiates between statistically indistinguishable solutions and solutions which are significantly different. Such a quantization of the space of structures can serve as a foundation of similarity based reasoning on structures to be extracted from data.

Publications

S.W. Kim and R.P.W. Duin, On Optimizing Dissimilarity-Based Classifier Using Multi-level Fusion Strategies (in Korean), *Journal of The Institute of Electronics Engineers of Korea*, Computer and Information (CI), vol. 45, no. 5, 2008, 15-24.

E. Pekalska and R.P.W. Duin, Beyond traditional kernels: classification in two dissimilarity-based representation spaces, *IEEE Transactions on Systems, Man Cybernetics*, vol. 38, no. 6, 2008, 729-744.

W.J. Lee and R.P.W. Duin, An Inexact Graph Comparison Approach in Joint Eigenspace, *Structural, Syntactic, and Statistical Pattern Recognition*, Proc. SSSPR2008 (Orlando, Florida, USA, 4-6 Dec 2008), Lecture Notes in Computer Science, vol. 5342, Springer Verlag, Berlin, 2008, 35-44.

R.P.W. Duin, E. Pekalska, A. Harol, W.J. Lee, and H. Bunke, On Euclidean corrections for non-Euclidean dissimilarities, Structural, Syntactic, and Statistical Pattern Recognition, Proc. SSSPR2008 (Orlando, Florida, USA, 4-6 Dec 2008), Lecture Notes in Computer Science, vol. 5342, Springer Verlag, Berlin, 2008, 551-561.

R.P.W. Duin and E. Pekalska, On refining dissimilarity matrices for an improved NN learning, *Proc. of the 19th Int. Conf. on Pattern Recognition (ICPR2008, Tampa, USA, Dec. 2008)*, IEEE Press, 2008.

B. Haasdonk and E. Pekalska, Indefinite Kernel Fisher Discriminant, *Proc. of the 19th Int. Conf. on Pattern Recognition (ICPR2008, Tampa, USA, Dec. 2008)*, IEEE Press, 2008.

M. Orozco-Alzate, R. P. W. Duin, and C. G. Castellanos-Dominguez, A generalization of dissimilarity representations using feature lines and feature planes, *Pattern Recognition Letters*, vol. 30, no. 3, 2009, 242-254.

M. Bicego, E. Pekalska, D.M.J. Tax, and R.P.W. Duin, Component-based Discriminative Classification for Hidden Markov Models, *Pattern Recognition*, 2009.

W.J. Lee and R.P.W. Duin, A Labelled Graph Based Multiple Classifier System, *MCS 2009*, accepted.

E. Pekalska, and B. Haasdonk, Kernel discriminant analysis with positive definite and indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2009, accepted).

Work package WP4:

Imposing geometricity on non-geometric similarities (embedding)

Work package leader: UNİYORK

Participants: UNİYORK, ETH Zurich, UNIVE, TUD

Start month: 7

End month: 30

Overall person-months: 62

The overall aim in this work package is that, given similarity data, possibly in the form of a weighted graph, we aim at developing algorithms for transforming them into instance-specific vectorial representations (embedding) that are suitable for traditional geometric learning algorithms.

Progress and Technical Achievements

Ihara Zeta Functions and Graph Vectorisation

We have explored a new structural characterisation of graphs based on the Ihara zeta function. Our motivation for embarking on this study was prior work at York, where we demonstrated that the zeta function is the moment generating function of the heat kernel trace and can be used to cluster graphs. Our study is based on the Ihara zeta function which is an extension of Riemann zeta function from prime numbers to prime cycles in a graph. The Ihara zeta function is constructed by first transforming a graph into its directed line-graph. It is the reciprocal of the characteristic polynomial of the transition matrix for the directed line-graph. Our characterisation of graphs is based on coefficients of the polynomial. The coefficients of the polynomial are related to the cycle frequencies in the graph, and are easily computed using the eigenvalues of the directed line-graph transition matrix (the so-called Frobenius operator). We have made a number of original developments to the study of the Ihara zeta function:

- a) We have shown how the Ihara coefficients can be used to cluster graph data-sets.
- b) We have shown how to generalise the Ihara zeta function to weighted graphs.
- c) We have started to explore how the Ihara zeta function can be generalised to hyper-graphs, and have obtained some initial experimental results on data-sets derived from image data.
- d) We have noted an interesting relationship between the Ihara zeta function and the discrete time quantum walk on a graph. The relationship is based on the observation that both rely on a representation based on the directed line-graph. In recent work, we have shown that the support matrix for the discrete time quantum walk can lift problems of cospectrality in trees and strongly regular graphs. We are currently investigating whether the Ihara zeta function also possesses this property.

Our future plans revolve around investigating whether the Ihara zeta function can be used to define a cycle kernel for graphs.

Heat Kernel Embeddings and Spectral Geometry

In prior work, we have shown how to associate curvatures with edges of a graph under the heat kernel embedding. The curvature is determined by the difference between the geodesic distance between nodes and the Euclidean distance between the embedded locations of the nodes. Our aim under SIMBAD has been to develop a deeper understanding of the geometric interpretation of the embedding, and to use this to develop a means of restoring metricity.

We have confined our attention to the embedding of triangulated weighted graphs, where the edge weights exponentially weight the similarity or affinity between nodes. We have made two steps in the towards the objectives of the work package.

- a) Our first step has been to show how the embedded triangles can be used to estimate Gaussian curvature using the Gauss-Bonnet theorem. The sets of curvatures extracted from a graph have been demonstrated to offer a means of computing graph similarity using the Hausdorff distance between sets, and clustering graphs.
- b) Our second advance has been to show how the Gaussian curvatures can be used to control the regularisation of graphs. The effect of applying this smoothing to the graph data has been to improve the cohesion of clusters that can be obtained.
- c) We are exploring alternatives to the heat-kernel embedding, which may provide a means for accounting for the node definiteness of the distance matrix. This is based on the wave equation, and allows to complex embeddings.

Our future plans are to develop a more sophisticated regularisation process which will improve the metricity of the embedding. The idea is to evolve the embedded triangles using a Ricci flow controlled by the estimated Gaussian curvatures. This can be reformulated as evolving the radii of hyperspheres on which the triangles are embedded.

Joint Eigenspaces

Stimulated by the visit of Pekalska and Duin to York the Delft researchers started to study new ways of representing sets of objects modelled by graphs. First a new way of comparing graphs was developed using joint eigenspaces. This resulted in an indefinite dissimilarity representation. The reasons behind that are studied in WP3. Using classifiers in the dissimilarity space, which does not suffer from the indefinite characteristic, good results have been found. In a later study the original graph comparison based on unattributed graphs was extended to graphs attributed with symbolic features. Based on a combining dissimilarities found for different symbols an entirely new approach to this problem was realised.

Another result of the Delft team was the insight that original indefinite representations may be transformed ("corrected") into less indefinite or fully definite ones. It was found that classifiers based on such corrected representations may occasionally improve but in other cases deteriorate the results. Further investigations are needed here.

Publications

- Bai Xiao, R. C. Wilson and E. R. Hancock, "Graph characteristics from the heat kernel trace", *Pattern Recognition*, doi:10.1016/j.patcog.2008.12.029, 2009.
- Bai Xiao, R. C. Wilson and E. R. Hancock, "A generative model for graph matching and embedding", *Computer Vision and Image Understanding*, doi:10.1016/j.cviu.2009.01.004, 2009.
- P. Ren, R. C. Wilson and E. R. Hancock, "Pattern vectors from the Ihara zeta function", DOI 10.1109/ICPR.2008.4761902, ICPR 2008.
- F. Escolano, E. R. Hancock, and M. A. Lozano, "Birkhoff polytopes, heat kernels and graph complexity", DOI 10.1109/ICPR.2008.4761921, ICPR 2008.
- S. Xia, P. Ren and E. R. Hancock, "Ranking the local invariant features for the robust visual saliencies", DOI 10.1109/ICPR.2008.4761170, ICPR 2008.
- Xia Shenping and E. R. Hancock, "3D object recognition using hypergraphs and ranked local invariant features", SSPR 2008, Lecture Notes in Computer Science, **5342**, pp. 117-126, 2008.
- Peng Ren, R. C. Wilson and E. R. Hancock, "Graph characteristics from the Ihara zeta function", SSPR 2008, Lecture Notes in Computer Science, 5342, pp. 256--266, 2008.
- Peng Ren, R. C. Wilson and E. R. Hancock, "Spectral embedding of feature hypergraphs", SSPR 2008, Lecture Notes in Computer Science, 5342, pp. 308--317, 2008.
- Xia Shenping and E. R. Hancock, "Clustering using class specific hypergraphs", SSPR 2008, Lecture Notes in Computer Science, 5342, pp. 318--328, 2008.
- H. ElGhawalby and E. R. Hancock, "Graph characteristic from the Gauss Bonnet Theorem", SSPR 2008, Lecture Notes in Computer Science, 5342, pp. 207-216, 2008.
- F. Escolano, M. A. Lozano and E. R. Hancock, "Polytopal graph complexity, matrix permanents, and embedding", SSPR 2008, Lecture Notes in Computer Science, 5342, pp. 237-246, 2008.
- H. ElGhawalby and E. R. Hancock, "Characterizing graphs using spherical triangles", IbPRIA 2009, to appear in Lecture Notes in Computer Science, 2009.
- Shenping Xia and E. R. Hancock, "Pairwise similarity propagation based graph clustering for scalable object indexing and retrieval", GbR 2009, to appear in Lecture Notes in Computer Science, 2009.
- H. ElGhawalby and E. R. Hancock, "Graph regularisation using gaussian curvature", GbR 2009, to appear in Lecture Notes in Computer Science, 2009.
- Peng Ren, R. C. Wilson and E. R. Hancock, "Characteristic polynomial analysis on matrix representations of graphs", GbR 2009, to appear in Lecture Notes in Computer Science, 2009.
- F. Escolano, D. Giorgi, E. R. Hancock, M. A. Lozano, and B. Falcidieno, "Flow complexity: Fast polytopal graph complexity and 3D object clustering", GbR 2009, to appear in Lecture Notes in Computer Science, 2009.

H. ElGhawalby and E. R. Hancock, "Geometric characterizations of graphs using heat kernel embeddings", Mathematics of Surface, to appear, Lecture Notes in Computer Science, 2009.

Shenping Xia and E. R. Hancock, "Learning class specific hypergraphs", ICIAP 2009, to appear, Lecture Notes in Computer Science, 2009.

D. H. White and R. C. Wilson, "Parts based generative models for graphs", ICPR 2008.

W. J. Lee and R. P. W. Duin, "An inexact graph comparison approach in joint eigenspace", SSSPR 2008, Lecture Notes in Computer Science, vol. 5342, Springer Verlag, Berlin, 2008, pp. 35-44.

W. J. Lee and R. P. W. Duin, "A labelled graph based multiple classifier system", MCS 2009, accepted for publication.

R. P. W. Duin, E. Pekalska, A. Harol, W. J. Lee, and H. Bunke, "On Euclidean corrections for non-Euclidean dissimilarities", SSSPR 2008, Lecture Notes in Computer Science, vol. 5342, Springer Verlag, Berlin, 2008, pp. 551-561.

R. P. W. Duin and E. Pekalska, "On refining dissimilarity matrices for an improved NN learning", ICPR 2008.

Work package WP5:

Learning with non-(geo)metric similarities

Work package leader: UNIVE

Participants: UNIVE, IST

Start month: 7

End month: 30

Overall person-months: 45

All approaches developed in WP2 and WP4 are based on the assumption that the non-geometricity of similarity information can be eliminated or somehow approximated away. When this is not the case, i.e., when there is significant information content in the non-geometricity of the data, alternative approaches are needed. The objective of this work package is to develop novel, general learning models which do not require the (geo)metric assumption, thereby working directly on the original data. Game theory offers an attractive and unexplored perspective that serves well our purpose.

The work package is divided into two main tasks: WP5.1 (“Study of equilibrium concepts”) and WP5.2 (“Generalizations”). In this report, we briefly describe the main achievements in the context of each of these tasks, and point to the relevant publications where these achievements are described in greater detail.

Progress towards the objectives of the Work Package

The aim of task WP5.1 was developing a game-theoretic framework based on a formalization of the competition between the hypotheses of class membership. According to this perspective, we shifted the focus from optima of objective functions to equilibria of (non-cooperative) games. We worked primarily on unsupervised learning problems. A preliminary attempt towards this general goal can be found in (Torsello, Rota Bulò, and Pelillo, 2006) where (a variation of) the concept of Nash equilibrium is proposed as a formalization of the notion of a cluster in a non-symmetric pairwise clustering context.

In the first six months of the work package we concentrated on extending the proposed approach to grouping and matching, developing new efficient algorithms for extracting Nash equilibria, generalizing the approach to high-order and contextual similarities, and exploring new equilibrium concepts applicable to machine learning problems. Details can be found in deliverable D5.1 and related publications.

Grouping and matching

We explored and extended the grouping framework introduced in (Torsello, Rota Bulò, and Pelillo, 2006). The first extension regarded the possibility of extracting overlapping clusters in a pairwise context, and it is based on two important properties of the game theoretical approach: First, the approach works as a multi figure/ground discrimination algorithm, extracting only cohesive groups, while leaving spurious entries unclustered.

Second, the clusters are extracted as surviving strategies at an equilibrium, thus different equilibrium can provide different, possibly overlapping, groups. Transforming the payoff matrix that drive the evolution of the selection process we can render unstable previously extracted equilibria, while not affecting any other cluster. The net result of this process is an approach to enumerate all possible groups approximately in order of relevance. This approach was used in (Rota Bulò, Torsello, Pelillo, 2009) to enumerate matches for shape recognition, while in (Torsello, Rota Bulò, Pelillo, 2008) the approach was generalized to continuous affinities and applied to perceptual grouping and image segmentation.

A further extension of the framework regarded the use of a game theoretic approach to matching and robust parameter estimation. In this framework matching can be formulated as a competition between correspondence hypotheses and the selection process leads to an equilibrium where only compatible correspondences survive. This matching process can then be used as in-lier selection for robust parameter estimation.

This idea was first explored in (Albarelli, Pelillo, and Viviani 2008) for an application to symmetry estimation and then further refined in (Albarelli, Rota Bulò, Torsello, and Pelillo, 2009).

Algorithms

We explored new efficient algorithms to extract Nash equilibria as a tool to achieve efficient classification. Building upon the invasion barrier paradigm, we proposed an Infection and Immunization Dynamics (InImDyn), modelling a plausible adaptation process in a large population. This dynamics exhibits a better asymptotic behaviour compared to other popular procedures like Fictitious Play and Replicator Dynamics, and can establish support separation in finite time, which can never be achieved by any interior-point method or any other evolutionary game dynamics. (Rota Bulò and Bomze, 2009). This last property is particularly interesting as it eliminates the need for an arbitrary threshold to extract the members of a cluster.

High order and contextual grouping

The game-theoretic framework naturally generalizes to allow k-way interactions among players, which is equivalent to using high-order similarity relations (hypergraphs). To follow this direction of investigation we generalized the Motzkin-Straus theorem relating cliques of a graph to the optima of a quadratic problem on the standard simplex, which is strongly related to the evolutionary stable strategies of our game theoretic clustering formulation. Our generalization links cliques of k-uniform hypergraphs to the minimizers of a polynomial optimization problem on the standard simplex (Rota Bulò and Pelillo, 2009). The problem is then optimized using a dynamical system that can be seen as a high-degree (or contextual) form of the classical replicator dynamics developed by Baum and Eagon in the late 1960's. This approach was used in (Rota Bulò, Albarelli, Pelillo, and Torsello, 2008) to achieve robust estimation of high-order parameters and generalized in (Rota Bulò and Pelillo, 2009) to clustering with continuous high-order affinities. Finally, in (Erdem and Torsello, 2009) we investigated the idea to of learning using contextual-dependent similarities. In particular, our game-theoretic approach was used to learn both the categories present in the data and the specific intra-category similarities that emerged from the context.

Equilibrium concepts for clustering

An initial investigation of classic game-theoretic concepts of equilibria showed that the vast majority of the method proposed in the literature is a refinement of the Nash equilibrium, where additional constraints added to offer stronger guarantees. Experiments on perceptual grouping and image segmentation clearly showed that the Nash equilibrium is already overly-restrictive, often leading to over-segmentation of the data. Motivated by this observation we decided to start our investigation on relaxations rather than refinements of the Nash equilibrium. To this end we developed the concept of *maximal good* which is defined as the set of strategies whose face is completely contained in the basin of attraction of an evolutionary stable strategy for all payoff-monotonic evolution dynamics, and provided an algorithm to compute it in the case of symmetric payoffs, based on the triangulation of the payoff matrix. Further, it can be shown that in the case of symmetric discrete 0-1 payoffs, i.e., the case described by the Motzkin-Straus theorem, the concept of cliques, dominant sets, and maximal good coincide.

References

- A. Torsello, S. Rota Bulò, and M. Pelillo (2006). Grouping with asymmetric affinities: A game-theoretic perspective. In *Proc. CVPR'06 - IEEE International Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, vol. 1, pp. 292-299.
- A. Torsello, S. Rota Bulò, and M. Pelillo (2008). Beyond partitions: Allowing overlapping groups in pairwise clustering. In *Proc. ICPR'08 - 19th International Conference on Pattern Recognition*, Tampa FL.
- S. Rota Bulò, A. Albarelli, M. Pelillo, and A. Torsello (2008). A hypergraph-based approach to affine parameter estimation. In *Proc. ICPR'08 - 19th International Conference on Pattern Recognition*, Tampa, FL.
- A. Albarelli, M. Pelillo, and S. Viviani (2008). Consensus Graphs for Symmetry Plane Estimation. In *SSPR'08 - Joint IAPR International Workshops on Structural, Syntactical, and Statistical Pattern Recognition*, pp. 197-206, Orlando, FL.
- S. Rota Bulò, A. Torsello, and M. Pelillo (2009). A game-theoretic approach to partial clique enumeration. *Image and Vision Computing* 27:911-922.
- S. Rota Bulò and M. Pelillo (2009). A generalization of the Motzkin-Straus theorem to hypergraphs. *Optimization Letters* 3(2):287-295.
- S. Rota Bulò and I. M. Bomze (2009). Infection and immunization: A new class of evolutionary game dynamics. *Games and Economic Behaviour* (submitted).
- A. Albarelli, S Rota Bulò, A. Torsello, and M. Pelillo (2009). Matching as a non-cooperative game. *ICCV'09 - IEEE International Conference on Computer Vision* (submitted).
- A. Erdem and A. Torsello (2009). A game-theoretic approach to jointly learn shape categories and contextual similarities (work in progress).
- S. Rota Bulò and M. Pelillo (2009). Clustering with higher-order similarities: A game-theoretic approach. *NIPS'09* (submitted)

Work package WP6:

Analysis of tissue micro-array (TMA) images of renal cell carcinoma

Work package leader: ETH Zurich

Participants: UNİYORK, UNIVE, UNIVR, IST, TUD

Start month: 13

End month: 36

Overall person-months: 48

The objective of this workpackage is to apply the techniques developed in workpackages WP2, WP4 and WP5 to the analysis of Tissue Micro Array (TMA) images of *renal cell carcinoma* (RCC), which is one of the ten most frequent malignancies in Western societies. Current diagnostic rules rely on exact counts of cancerous cell nuclei that are manually counted by pathologists.

Originally, the ETH Zurich group has proposed to provide a data set of tissue micro-arrays from renal clear cell carcinoma biopsies as a test data set for similarity based pattern recognition. The progression state of cancer is characterized by the number of cancer cells which divide at any point in time. Therefore, it is important to separate these cells from normal proliferating cells which are also stained by a proliferation marker. The pathologists claim to detect cancer cells based on the shape of the cell nuclei. The ETH group currently prepares a curated data set of images with cancer cells and normal cells such that different methods can be tested on this benchmark problem.

Since January 2009, a new collaboration with the group of Prof. Krek has been started to analyze prostate cancer data for biomarker detection. The prostate cancer data are composed of microarray data, mass spectrometry data and tissue microarrays. We are currently in close contact with the biologists to identify the biological hypotheses which we will test in the framework of the SIMBAD project. We consider this prostate cancer study as an excellent opportunity for the SIMBAD project since these data are unique in quality, in their heterogeneity and in their relevance for medical research. Therefore, we have decided to use personal resources of WP6 earlier than planned to respond to this demand.

Work package WP7:

Analysis of brain magnetic resonance (MR) scans for the diagnosis of mental illness

Work package leader: UNIVR

Participants: UNIVR, UNIVE, IST, ETH Zurich, TUD

Start month: 13 End month: 36 Overall person-months: 51

The objective of work package 7 (WP7) is to apply similarity-based techniques and algorithms developed in the other work packages to the analysis of brain magnetic resonance (MR) images in the context of mental health research (e.g., schizophrenia). Data consists of morphological MR images (3DA), instrumental to exploring the content of grey and white matter tissues, the volume of specific structures and the 3D shape morphology of particular brain regions, and diffusion weighted imaging (DWI) images, providing information on the microstructural integrity of the brain.

In order to acquire knowledge, experience and data regarding brain MR imaging and to release a first dataset to the SIMBAD partners, the start of WP7 has been anticipated by 5 months with respect to the scheduling programmed in the Annex.

Summary of progress

Our medical partners involved in the Verona-Udine brain imaging and neuropsychology program (VUBINP), operating one of the largest worldwide MR databases for schizophrenia, have provided the SIMBAD project with several raw and processed data that have been thoroughly checked and organized.

UNIVR has developed a naming scheme to give an overall characterization of the data types present in the database, starting from raw data at lower levels of the scheme to processed data at the higher levels. This scheme highlights the intermediate steps taken by established medical procedures to perform analytical experiments:

- raw morphological images are manually re-aligned to compensate for head pose variations during data acquisition, and are re-sampled to a standard format;
- *tracing* is an operation performed by a trained expert with the goal of isolating a particular region of interest (ROI), such as amygdala, hippocampus, thalamus or intracranial volume (ICV);
- in the case of DWI images, tracing is substituted by sampling through circular regions of the apparent diffusion coefficient (ADC) values;
- statistical analyses are performed on the volumetric and ADC measurements, using ICV and clinical data as covariates.

The overall goal in medical experiments is to differentiate the population of schizophrenic patients from the normal controls in such a way as to highlight physiological abnormalities that can explain behavioral and

cognitive aberrations in patients. UNIVR has acquired knowledge and expertise about the procedures and the data to guarantee the correctness and accuracy of the data provided to the SIMBAD partners.

The database currently managed by UNIVR contains 150 subjects, with 71 schizophrenic patients and 79 normal controls. Briefly: at level 1, there are 150 3DA scans, 143 transversal DWI scans and 140 cortical DWI scans; at level 2, there are 150 re-sampled re-aligned images; at level 3, there are 7 ROIs (amygdala, dorsolateral prefrontal cortex, entorhinal cortex, Heschl's gyrus, hippocampus, superior temporal gyrus, thalamus) plus the ICV. Clinical data contains information about age, gender, first diagnosis, years of illness, education, profession, weight, height, smoking habits, alcohol use, and drugs use.

From this database, UNIVR has identified a first benchmark dataset for the application of similarity-based classification techniques. This dataset contains volume data about the 7 traced ROIs. These ROIs have been identified by our medical partners and by the medical community at large as functional parts that have a definite role in schizophrenia and, therefore, are promising for developing diagnosis tests for this illness. This dataset is currently downloadable by SIMBAD partners from the restricted area of the SIMBAD website, and freely usable for research purposes after signing an agreement devised by our partners in VUBINP.

By releasing this dataset, UNIVR pursues the SIMBAD objective of providing a challenging testbed with important research significance for the validation of similarity-based methods and algorithms. Regarding the challenging aspect of the given data, UNIVR has performed statistical analyses of the ROIs volumetric data, following the prescribed medical guidelines and standards, and confirmed that some ROIs (entorhinal cortex and dorsolateral prefrontal cortex) show significant abnormalities in the population of schizophrenic patients. This is in accordance with findings in the medical community.

However, the use of volumetric data alone is not powerful enough to discriminate between patients and controls with an effective classification rate. UNIVR has initiated a series of experiments based on the analysis of histograms of values present within the ROIs, corresponding roughly to evaluating the different distribution of grey matter, white matter, or a mixture of both in the two populations. In the process of performing such experiments, UNIVR has identified additional procedures and recommendations for the pre-processing of the data (included in a technical report to be published within SIMBAD).

Preliminary tests performed with a variety of feature extraction methods (e.g., quantization, PCA) and off-the-shelf classifiers (e.g., svm, knn) show promising results (all the details may be found in [1]). Three ROIs have been found especially discriminative: amygdala, hippocampus and dorsolateral prefrontal cortex, in particular on the left side. The abnormalities in these regions are consistent with findings in MRI studies in schizophrenia. This suggests that the image content of the ROIs, beyond the volumetric evidence, might offer enough information to effectively classify schizophrenic patients. With this task, UNIVR proposes to establish a baseline of pre-processing methods and classification results on the given dataset as a comparative reference for future similarity-based experiments.

References

[1] D. S. Cheng, M. Bicego, U. Castellani, S. Cerruti, M. Bellani, G. Rambaldelli, M. Aztori, P. Brambilla, V. Murino: "Schizophrenia Classification Using Regions of Interest in Brain MRI", submitted to INTELLIGENT

DATA ANALYSIS IN BIOMEDICINE AND PHARMACOLOGY, in conjunction with the 12th Conference on Artificial Intelligence in Medicine (AIME) 2009.

Work package WP8:

Dissemination, communication and exploitation

Work package leader: UNIVE

Participants: UNIVE, IST

Start month: 7

End month: 36

Overall person-months: 7

Objective of this Work package is to ensure that the results achieved within the project was publicly available and to disseminate them in the most suitable way, not only at the scientific community level

Progress towards the objectives of the Work Package

The consortium is ensuring the highest diffusion of the research results, both inside and outside the consortium.

In order to provide a timely access of information within the consortium, and increase interactions among the SIMBAD partners, we established the *SIMBAD Technical Report Series*: all partners are asked to publish their latest results, as soon as they are produced, in the form of a Technical Report which will be then made available only to the other partners in the restricted area of the project website.

All partners are periodically informed on the latest publications in the series by e-mail.

The dissemination of the project results took place mainly through presentations at the leading international conferences and workshops and in few cases publications in the specialized technical journals.

The entire list of publications is available in the Annex – List of publications, publications referred to each work package are also listed at the end of the work packages described above.

The bibliographic references are publicly available via the project website (<http://simbad-fp7.eu/bibliography.php>).

We aim at establishing a series of workshops specifically devoted to the project theme. In the hope that the topics covered within the SIMBAD project will have a lasting and substantial impact within scientific community, the workshops will possibly continue even after the end of the project activities.

We are planning to run the first edition of the workshop “*SIMBAD 2010*” that will take place in Venice (Italy), early in 2010.

A dedicated web site (<http://simbad-fp7.eu>) has been established (see section 5 for details)

The activities related to the SIMBAD project have been covered on the press via several interviews which are available at the project website (<http://simbad-fp7.eu/news.php>)

We have also produced a thousand copies of the following leaflet:

Starting date: 01 April 2008
Duration: 36 months

<http://simbad-fp7.eu>
contact: info@simbad-fp7.eu


 Università Ca' Foscari Venezia - Italy
<http://www.dsi.unive.it>
 Marcello Pelillo - pellillo@dsi.unive.it


 COMMISSION OF THE EUROPEAN COMMUNITIES


 University of York - United Kingdom
<http://www.york.ac.uk>
 Edwin Robert Hancock - erh@cs.york.ac.uk

INFORMATION SOCIETY AND MEDIA
DIRECTORATE-GENERAL

 Seventh Framework Programme
Information and Communication Technologies

 Collaborative Project
FET Open


 Technische Universiteit Delft - The Netherlands
<http://fict.ewi.tudelft.nl/>
 Robert Duin - r.duin@ieee.org


Beyond Features
Similarity-based pattern analysis
and recognition


 Instituto Superior Técnico - Portugal
<http://www.ist.utl.pt>
 Mário Figueiredo - mario.figueiredo@lx.it.pt


 Università degli Studi Di Verona - Italy
<http://vips.sci.univr.it>
 Vittorio Murino - vittorio.murino@univr.it


 Eidgenössische Technische
Hochschule Zürich - Switzerland
<http://www.ethz.ch>
 Joachim Buhmann - jbuhmann@inf.ethz.ch



BEYOND FEATURES	WHY SIMBAD?	SIMBAD AT A GLANCE
<p>Traditional pattern recognition techniques are centered on the notion of "feature". According to this view, each object is described in terms of a vector of numerical attributes and is therefore mapped to a point in a Euclidean (geometric) vector space so that the distances between the points reflect the observed (dis)similarities between the respective objects.</p> <p>Despite its potential, the geometric approach suffers from a major intrinsic limitation, which concerns the representational power of feature-based descriptions. In fact, there are numerous application domains where either it is not possible to find satisfactory features or they are inefficient for learning purposes. Most commonly, this is typically the case when objects are described in terms of structural properties, such as parts and relations between parts, as is the case in shape recognition.</p> <p>This project aims at bringing to full maturation a paradigm shift that is currently just emerging within the pattern recognition and machine learning domains, where researchers are becoming increasingly aware of the importance of similarity information per se, as opposed to the classical feature-based approach. Indeed, the notion of similarity (which appears under different names such as proximity, resemblance, and psychological distance) has long been recognized to lie at the very heart of human cognitive processes and can be considered as a connection between perception and higher-level knowledge, a crucial factor in the process of human recognition and categorization.</p>	<p>By departing from vector-space representations one is confronted with the challenging problem of dealing with (dis)similarities that do not necessarily possess the Euclidean behavior or not even obey the requirements of a metric. The lack of the Euclidean and/or metric properties undermines the very foundations of traditional pattern recognition theories and algorithms, and poses totally new theoretical/computational questions and challenges.</p> <p>We aim at undertaking a thorough study of several aspects of similarity-based pattern analysis and recognition methods, from the theoretical, computational, and applicative perspective, with a view to substantially advance the state of the art in the field, and contribute towards the long-term goal of organizing this emerging field into a more coherent whole.</p> <p>The whole project will revolve around two main themes, which basically correspond to the two fundamental questions that arise when abandoning the realm of feature-based representations:</p> <ol style="list-style-type: none"> 1. How can one obtain suitable similarity information from object representations that are more powerful than, or simply different from, the vectorial? 2. How can one use similarity information in order to perform learning and classification tasks? <p>According to this perspective, the very notion of similarity becomes the pivot of non-vectorial pattern recognition in the same way as the notion of feature-vector plays the role of the pivot in the classical (geometric) paradigm.</p>	<p>From a methodological perspective, SIMBAD will be structured around the following strands:</p> <p>Deriving similarities for non-vectorial data (structural kernels), to develop computational models for generating similarities for non-vectorial data, with particular emphasis on structured and semi-structured descriptions</p> <p>Foundations of non-(geo)metric similarities, to study both the causes of the lack of (geo)metricity in the similarity data and its effects over traditional machine learning algorithms</p> <p>Imposing geometricity on non-geometric similarities (embedding), to develop algorithms for transforming the original similarity data into proper vectorial representations suitable for traditional learning algorithms</p> <p>Learning with non-(geo)metric similarities, to develop novel, general learning models which do not require the (geo)metric assumption</p> <p>An important part of SIMBAD will concern the validation of the developed techniques, focusing mainly on biomedical problems.</p> <p>Analysis of tissue micro-array (TMA) images of renal cell carcinoma, to validate the techniques developed by applying them to the analysis of Tissue Micro Array (TMA) images of renal cell carcinoma</p> <p>Analysis of brain magnetic resonance (MR) scans for the diagnosis of mental illness, to validate the techniques developed by applying them to the analysis of brain MR scans in the context of mental health research</p>



Personnel

UNIVE: Assigned to the SIMBAD project are Marcello Pelillo (associate professor), heading the UNIVE contribution, Andrea Torsello (assistant professor) and, since November 2008, Aykut Erdem (post-doc). As a result of a previous cooperation on the topics of SIMBAD, Samuel Rota Bulò (postdoc at the University of Venice) is also actively participating in the project. Veronica Giove, project administrator, assisting the project coordinator and Sonia Barizza, the Administrative Responsible of the Computer Science Department, deal with management activities .

UNİYORK: The overall project plan has been to initially progress the project with established research students in their second and third years, and then to train-up two new students who can start to make technical progress in their second and third years. The established research students are Peng Ren, Howaida El Ghawalby and David White. On October 1st 2009 we recruited two new students Lin Han and Eliza Xu. Lin has a masters in Informatics from Edinburgh and Eliza a masters in Machine Learning from Bristol. Both are Chinese nationals. Han will work on generative models and Xu on embedding methods and non-definite kernels.

TUD: Assigned to the SIMBAD project are Wan-Jui Lee (100%, post-doc) from the start of the project and Robert P.W. Duin (partially, associated professor), heading the Delft contribution. As a result of a historical cooperation also Elzbieta Pekalska (research associate at the University of Manchester, UK) is actively participating in the project. Starting in September 2008 Alessandro Ibba is working as a PhD student in the field of dissimilarity based pattern recognition as well. He is not payed from the project. Finally, Marco Loog (assistant professor) and Marcel J.T. Reinders (full professor) have been following the project and may attend some meetings. Wan-Jui Lee enjoyed maternity leave from October 2008 until February 2009. Since then she has been further working for 80% of time.

IST: Assigned to the SIMBAD project are Mario Figueiredo, Ana Fred, Pedro Aguiar (faculty members) and André Martins, David Pereira-Coutinho, and Arthur Ferreira (PhD students)

UNIVR: Assigned to the SIMBAD project are Vittorio Murino, Umberto Castellani, Manuele Bicego, Marco Cristani (faculty members), Dong Seon Cheng (post-doc), Anna Carli, Alessandro perina, and Marcella Bellani (PhD students).

ETH Zurich: During the first year the two PhD students Sharon Wulf (since Nov. 15, 2008) and Peter Schüffler (since Dec. 1, 2008) worked on the SIMBAD project in the ETH Zurich group. This work has been jointly supervised by Volker Roth, who is a professor at the University of Basel, Dr. Cheng Soon Ong,

research associate at ETH Zurich and Prof. Joachim M. Buhmann. It should be noted that the search for qualified PhD students required half a year.

4. Deliverables and milestones tables

TABLE 1. DELIVERABLES									
Del. no.	Deliverable name	WP no.	Lead beneficiary	Nature	Dissemination level	Delivery date from Annex I (proj month)	Delivered Yes/No	Actual / Forecast delivery date	Comments
D 1.1	Quality Plan	1	1	R	PU	1	Yes		
D 8.3	Project Leaflet	8	1	O	PU	7	Yes		
D 8.2	First Dissemination Plan	8	1	R	PU	7	Yes		
D 8.1	Project Web Page	8	1	O	PU	7	Yes		
D 2.1	Compression Kernels	2	4	R	PU	12	Yes		
D 5.1	Equilibrium concepts for pattern recognition	5	1	R	PU	12	Yes		
D	Study on	3	3	R	PU	12	Yes		

3.1	(non)geometricity								
D 3.2	Characterization of invariances	3	6	R	PU	12	Yes		

Milestones

TABLE 2. MILESTONES							
Milestone no.	Milestone name	Work package no	Lead beneficiary	Delivery date from Annex I	Achieved Yes/No	Actual / Forecast achievement date	Comments
MS 1	Equilibrium Concepts	5	1	12	Partially		We have only partially achieved this milestone. (see section 5 of this deliverable) There will be no impact on the other parts of the project, though. We plan to achieve this milestone early in the second year of the project
MS 2	Foundations	3	3	12	Yes		

5. *Project management*

Management tasks and achievements

Financial distribution: University Ca' Foscari of Venice (UNIVE) acting as the Coordinator, as a first step, distributed the Community Financial Contribution among the beneficiaries, following the allocation reported in the Contract, without any delay.

Bank account: Even though the Computer Science Department keeps one single bank account, it is possible, at any time, to determine the actual balance and the flow of money and the quarterly return of interests gained on the prefinancing coming from European Commission funds.

Consortium Agreement: UNIVE, in accordance and together with the other partners, provided the Consortium Agreement. This document aims at covering the interests of each partner and at protecting their rights.

Monitoring the project implementation: UNIVE is monitoring the implementation of the project, looking at the compliance of the partners obligations: activities progress, sharing of the results, settlement of the working groups, deliverables (quality plan, project leaflet, first dissemination plan, project web site designing).

Communication of data for evaluation: UNIVE together with all the SIMBAD partners collaborated with the DG Information Society & Media to assess the progress in 2008 towards the achievement of the IST-RTD implementation objectives, gathering the results of the SIMBAD project.

Problems which have occurred

No problem arose from the management of the consortium. All discussions have been carried out during the project meeting, and/or via e-mail.

Changes in the consortium

No amendment of the Consortium has been necessary

List of project meetings, dates and venues

Project Kick off meeting: the first meeting of the project took place in Venice (IT), at the University of Ca' Foscari, on the 18th April 2008.

Project Intermediate meeting: The second meeting was held in York (UK) on the 14th and 15th of November 2008.

Project planning and status

The work carried out between months 1- 12 in Work Packages WP3, WP4, WP8 is in line with Annex I.

As concerns WP2, Task 2.1, UNIVR expects to need more person-months in this task in order to complete the work carried out during the first year. Unfortunately at this stage it is difficult to precisely quantify the number of extra person-months. It would also be possible that it will be required a shift in the budget figures, in order to re-equilibrate the whole budget and to complete the assigned work. Also in this case, it is now hard to explicitly quantify the possible variation in the budget figures.

As concerns WP2, task 2.2, as reported in Deliverable D2.1, IST did some progress towards "compression kernels", although not as much as anticipated. IST implemented and tested techniques based on Lempel-Ziv and Burrows-Wheeler text compression algorithms, which can be used to estimate: the entropy of a source assumed to have generated a given sequence of symbols; the Kullback-Leibler divergence between two sources assumed to have generated a given pair of sequences of symbols. These are the fundamental building blocks of "compression kernels" for sequences of symbols (text). What we have not done yet is use these compression-based estimators in a kernel-based learning method, such as a support vector machine. There is no actual "deviation" from what we have planned, but simply some delay. We plan to finish our work in the first few months of the second year of the project. There will be no impact on the rest of the project.

As concerns WP 5, task 5.1, as reported in deliverable D5.1, an initial investigation of classic game-theoretic concepts of equilibria showed that the vast majority of the method proposed in the literature is a refinement of the Nash equilibrium, where additional constraints are added to offer stronger guarantees. Experiments on perceptual grouping and image segmentation clearly showed that the Nash equilibrium is already overly-restrictive, often leading to over-segmentation of the data. Motivated by this observation we decided to start our investigation on relaxations rather than refinements of the Nash equilibrium, and to postpone to the second year of the project a deeper study on classical equilibrium concepts, when a new post-doc will be also available. Again, there will be no impact on the rest of the project

Concerning WP6, ETH Zurich used personal resources for WP6, earlier than planned: since January 2009, a new collaboration with the group of Prof. Krek has been started to analyze prostate cancer data for biomarker detection. The prostate cancer data are composed of microarray data, mass spectrometry data and tissue microarrays. ETH Zurich is currently in close contact with the biologists to identify the biological hypotheses which will be tested in the framework of the SIMBAD project. This prostate cancer study as an excellent opportunity for the SIMBAD project since these data are unique in quality, in their heterogeneity and in their relevance for medical research.

Concerning WP7, in order to acquire knowledge, experience and data regarding brain MR imaging and to release a first dataset to the SIMBAD partners, UNIVR started WP7 (Analysis of brain magnetic resonance (MR) scans for the diagnosis of mental illness) 5 months before the scheduling programmed in Annex I.

UNIVR expects to need more person-months in Work packages WP2 and WP7 in order to complete the assigned work.

UNIVR spent up to now 22 out of the 56 person months foreseen: during this first 12 months UNIVR worked on WP7 to prepare and organize data in order to make them available at the end of the first year to the other SIMBAD partners. Furthermore, there are some additional person months necessary to carry out the activities related to WP 2.1. Actually, the activities related to UNIVR were underestimated during the preparation of the proposal.

It is now possible to quantify the number of extra person months foreseen to 8, for a total number of 64 person month.

The new distribution of person month is reported on the following table:

UNIVR	WP 2.1	WP 2.2	WP 6	WP 7	WP 8	Total
foreseen PM	18	10	7	20	1	56
updated PM	22	10	7	24	1	64

The additional man months correspond to 27.300 €. This amount will be taken from the "Other direct costs" figure. The updated budget of UNIVR is detailed in the table here below:

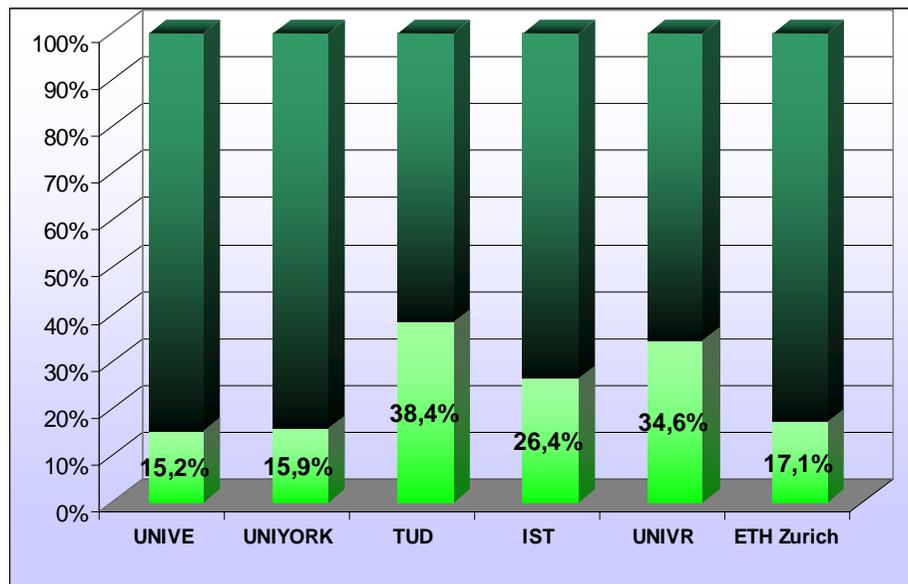
	RTD / Innovation Foreseen figures	RTD / Innovation Updated figures
Personnel costs	€ 170.500,00	€ 197.800,00
Subcontracting	€ 0,00	€ 0,00
Other direct costs	€ 60.000,00	€ 32.700,00
Indirect costs	€ 138.300,00	€ 138.300,00
Total costs	€ 368.800,00	€ 368.800,00
Requested EC contribution	€ 276.600,00	€ 276.600,00

Comparison between actual person months allocated in the first 12 months and the total planned resources

	WP 1	WP 1	WP 2.1	WP 2.1	WP 2.2	WP 2.2	WP 2.3	WP 2.3	WP 3.1	WP 3.1	WP 3.2	WP 3.2	WP 4.1	WP 4.1	WP 4.2	WP 4.2	WP 5.1	WP 5.1	WP 5.2	WP 5.2	WP 6	WP 6	WP 7	WP 7	WP 8	WP 8	TOTAL PM	TOTAL PM	
UNIVE	4,8	12,0						5,0					5,0				3,0	15,0	4,1	25,0		6,0		6,0		2,0	11,9	76,0	
UNİYORK						2,0							12,5	40,0		2,0						11,0		11,0		1,0	12,5	67,0	
TUD			1,0						11,0	12,0	1,4	6,0	2,0	2,0		2,0							4,0		4,0		1,0	15,4	31,0
IST				5,0	9,2	5,0		5,0												5,0		5,0		5,0		1,0	9,2	31,0	
UNIVR			16,0	18,0	2,0	10,0																7,0	4,0	20,0		1,0	22,0	56,0	
ETH Zurich						2,0			2,0	3,0	2,5	8,0		2,0		9,0						4,0	15,0		5,0		1,0	8,5	45,0
tot	4,8	12,0	17,0	23,0	11,2	19,0	0,0	10,0	13,0	15,0	3,9	14,0	14,5	49,0	0,0	13,0	3,0	15,0	4,1	30,0	4,0	48,0	4,0	51,0	0,0	7,0	79,5	306,0	

Actual Planned

Comparison between actual expenses and budget of the action



Activities foreseen in the project, in the near future are the following:

1. Third project meeting, scheduled for the 26th and 27th June 2009, in Zurich.
2. First edition of the workshop “SIMBAD 2010”: it that will take place in Venice, Italy, early in 2010 (see the previous section n. 6 Work progress and achievements during the period (months 1–12), in the Work package WP8 description. A more detailed programme of the event will be developed during the Zurich project meeting)

Impact of possible deviations from the planned milestones and deliverables, if any

The minor deviation from the planned milestone MS 1 “Equilibrium Concepts” will have no significant impact on the rest of the project.

All other deliverables and milestones foreseen in the project have been achieved.

Changes to the legal status of any of the beneficiaries

No change in the legal status of the beneficiaries is to be reported

Development of the Project website

A dedicated web site has been established. In the website project publications, reports, software, data, and results, both theoretical and technological, will be published and illustrated. An important part of the site will be the collection of similarity data that will be publicly available, thereby creating a common reference for similarity-based computational models and algorithms. simbad-fp7.eu serves as a knowledge-base for similarity-based approaches to pattern recognition and machine learning.

Information about project development are updated by each consortium member thanks to his own restricted web site area.

After the project overview in the homepage, there are several sections:

1. **Partners descriptions**, name and address of the scientific responsible for the unit.
2. Main **deadlines** to be respected
3. Useful downloadable **documents** : i.e. Guidance Notes on Project Reporting, Guide to Financial Issues relating to FP7 Indirect Actions, some notes on travels to be charged to the project budget, SIMBAD Leaflet, SIMBAD Consortium Agreement.
4. **Deliverables**: here is the deliverables list and the downloadable deliverables, when available
5. Past and present **events**: this is the place where main events strictly related to the project (project meetings) or related to the scientific community events are reported
6. **Bibliography**: bibliographic references
7. **Work packages list**

8. **Press Coverage:** list of interviews, where the activities related to the SIMBAD project have been advertised
9. **Restricted Area:** area where confidential level documents are stored (e.g. software, data, and results)

Use of foreground and dissemination activities during this period

Participation in International Conferences: the first positive results of the project activities have been disseminated within the scientific community through the participation of some of the members of the consortium in three International conferences:

1. The 12th IAPR International Workshop on Structural and Syntactic Pattern Recognition, Orlando (SSPR), Florida (USA).
2. ICPR 2008 International Conference on Pattern Recognition, Tampa, Florida (USA)
3. International Conference on Image Analysis and Recognition (ICIAR) 2008, June 2008, Portugal

Several people participated in the first two conferences (SSPR and ICPR). The coordinator, Marcello Pelillo, requested (and obtained) to the Project Officer the agreement to attend to this conference for the following people:

From UNIVE: Marcello Pelillo, Andrea Torsello, Andrea Albarelli, to present three podium presentations, one poster presentation

From UNIYORK: Edwin Hancock submitted SIMBAD results: there were six podium presentations and three poster presentations.

From TUD: Wan-Jui Lee with two presentations

Samuel Rota Bulò, participated in Doctoral Workshop on Game Theory during October 2008 at the University of Konstanz.

List of publications:

In the first year the Simbad partners produced 51 publications related to the project topics. In the “Annex A – SIMBAD list of publications”, we listed only the publications carried out thanks to the SIMBAD funds (which include proper acknowledgements).

The list of publications is available on the project webpage as well (<http://simbad-fp7.eu/bibliography.php>).

Leaflet: We have also produced a thousand copies of a leaflet describing the project’s activities (see section 3, WP8 for details)

Press Coverage: The activities related to the SIMBAD project have been advertised via several interviews which are available at the project website (<http://simbad-fp7.eu/news.php>).

6. Explanation of the use of the resources

TABLE 3.1 PERSONNEL, SUBCONTRACTING AND OTHER MAJOR DIRECT COST ITEMS FOR BENEFICIARY 1 UNIVERSITÀ CA' FOSCARI DI VENEZIA FOR THE PERIOD FROM 01/04/2008 TO 31/03/2009			
Work Package	Item description	Amount	Explanations
1	Personnel costs	13.845 €	Salaries of: Assistant Professor for 0,4 months administrative responsible for 0,7 months recruited collaborator for 4,7 months
2, 4, 5	Personnel costs	20.950 €	Salaries of: two Assistant Professors for 3,10 months one recruited fellowship for 4 months
2,4,5	Travel costs	10169 €	participation in project York Meeting, research activities in York, participation in "International conference on Pattern Recognition", and "Structural and Syntactic Pattern Recognition", FLORIDA (USA), participation in workshop in Costanza (Germany), in seminar at the University of Vienna, in workshop Learning Intelligent Optimization LION 3
2, 4, 5	Equipment Purchase	448 €	Purchase of laptop, desktop computer to implement the specif research
1	Remaining direct costs	378 €	Minor items
2, 4, 5	Remaining direct costs	1.749 €	Minor items
TOTAL DIRECT COSTS		47.540 €	

Personnel

Marcello Pelillo, associate professor, heading the UNIVE contribution was involved for 0,4 month in the Management activities and for 1,4 PM in the RTD activities.

Andrea Torsello, assistant professor, worked for 1,7 PM on the project. Aykut Ibrahim Erdem, started with his post-doc grant in November 2008 and he is 100% charged on the project.

Veronica Giove, project administrator, worked for 3,7 PM on the project.

Sonia Barizza is the Administrative Responsible charged for 0,7 PM on the project

Prof. Immanuel Bomze from the University of Vienna, has visited UNIVE partner to do work related to WP 5 on September 2008.

Travels

Marcello Pelillo, Andrea Torsello and Andrea Albarelli submitted the SIMBAD results to ICPR'08: the 19th International Conference on Pattern Recognition, and SSPR workshop (the 12th IAPR International Workshop on Structural and Syntactic Pattern Recognition), in Orlando, FL

(<http://ml.eecs.ucf.edu/ssspr/sspr.php>) since they got three podium presentations and one poster presentation. The agreement of the European Officer received in December 2008, allowed the participation in the Conferences.

Marcello Pelillo, Andrea Torsello, Samuele Rota Bulò and Veronica Giove participated in the York project Meeting.

Andrea Torsello, made some research activities in York.

Samuel Rota Bulò participated in workshop in Costanza (Germany), and travelled for a short visit at the University of Vienna, participated in workshop Learning Intelligent Optimization LION 3 (Trento, Italy)

Report on additional resources

Samuel Rota Bulò (4 PM) and Andrea Albarelli (2 PM) are post-doc not paid with SIMBAD funds, but involved in the research activities.

TABLE 3.1 PERSONNEL, SUBCONTRACTING AND OTHER MAJOR DIRECT COST ITEMS FOR BENEFICIARY 2, UNIVERSITY OF YORK, FOR THE PERIOD FROM 01/04/2008 TO 31/03/2009			
Work Package	Item description	Amount	Explanations
4.1 – 4.2	Personnel costs	15.344,88 €*	Salaries of one full professor, two PhD students
4.1 – 4.2	Travel	11.049 €*	Travel Costs: travel to Venice (Italy), participation in SSPR 2008, ICPR 2008 and GbR 2009 in the USA
	Other costs	3.927 €	Equipment: 2 laptop purchase
TOTAL DIRECT COSTS		30.320,83 €*	

Personnel

Professor Edwin Hancock, he is full professor leading York strand and WP4, he is charged for the 0,5 PM.

Lin Han, Since 1/10/08 PhD student, she worked for the project for 6 PM

Eliza Xu, Since 1/10/08 PhD student, she worked for the project for 6 PM

Travels

Edwin Hancock and Richard Wilson participated in the Venice project kick-off meeting.

Edwin Hancock travelled to the USA to present at ICPR'08 and SSPR workshop six podium presentations and three poster presentation.

York hosted the project meeting in November 2008. The costs of venue hire (St William's College) and catering were charged to the project.

Report on additional resources

The following people were involved in the SIMBAD project research, but have not been paid with SIMBAD funds: Dr Richard Wilson, Reader; Peng Ren, PhD student; Howaida El Ghawalby, PhD student; David White, PhD student

TABLE 3.1 PERSONNEL, SUBCONTRACTING AND OTHER MAJOR DIRECT COST ITEMS FOR BENEFICIARY 3, TECHNISCHE UNIVERSITEIT DELFT, FOR THE PERIOD FROM 01/04/2008 TO 31/03/2009			
Work Package	Item description	Amount	Explanations
2, 3	Personnel costs	69.568,48 €	<i>Salaries for one researcher for 12 month and one Associate Professor for 3,4 PM</i>
2,3	Travel Costs	5.098,19 €*	<i>United Kingdom, USA</i>
TOTAL DIRECT COSTS		74.666,67 €*	

Personnel

Robert Duin, Associate Professor worked on the project for 3,4 PM and 12 PM Wan-Jui Lee (PhD student).

Travels

The Unit used SIMBAD project funds for the visit to York (UK) for scientific exchange, for the Venice kick-off meeting, for the York project meeting (14 and 15 November, 2008), for the paper presentations for two conferences in USA: The 12th IAPR International Workshop on Structural and Syntactic Pattern Recognition, Orlando (SSPR), Florida (USA) and the ICPR 2008 International Conference on Pattern Recognition, Tampa, Florida (USA). The Project Officer allowed the participation in both the Conferences on December 2008 due to the need to present their posters.

Report on additional resources

Elzbieta Pekalska (research associate at the University of Manchester, UK) contributed to some SIMBAD papers and deliverables. She also attended the Venice project meeting. An estimation of her work for the project is 200 hours , Marco Loog (assistant professor) worked for the project for 80 hours, and Marcel J.T. Reinders (full professor) gave their contribute with around 10 hours to the scientific achievement of the project, even though they are not paid with SIMBAD funds.

TABLE 3.1 PERSONNEL, SUBCONTRACTING AND OTHER MAJOR DIRECT COST ITEMS FOR BENEFICIARY 4, INSTITUTO SUPERIOR TECNICO, FOR THE PERIOD FROM 01/04/2008 TO 31/03/2009			
Work Package	Item description	Amount	Explanations
2	Personnel costs	44.887,17 €	<i>Salaries of 3 researchers</i>
2	Subcontracting	0€	
2	Travel	3.533,13 €*	<i>Project meeting UK</i>
2	Other Direct Costs	0€	
	Remaining direct costs	0€	
TOTAL DIRECT COSTS		48.420,30 €	

Personnel costs

Mario Figueiredo (Associate Professor) worked on the project for 3,4 PM, Ana Fred (Assistant Professor) for 3,14 PM and Pedro Aguiar (Assistant Professor) for 3,02 PM.

Travel

Expenditures of this Unit are only related to the participation in the Project Meeting in York, 14 and 15 November 2008.

Report on additional resources

IST contributed with its own human resources to the project:

André Martins, David Pereira-Coutinho, Arthur Ferreira are PhD students that have been involved in the SIMBAD activities.

TABLE 3.1 PERSONNEL, SUBCONTRACTING AND OTHER MAJOR DIRECT COST ITEMS FOR BENEFICIARY 5, UNIVERSITÀ DEGLI STUDI DI VERONA FOR THE PERIOD FROM 01/04/2008 TO 31/03/2009			
Work Package	Item description	Amount	Explanations
2,7	Personnel costs	76.063,30 €	<i>Salaries of 1 postdoc for 4 months, 1 postdoc for 3 months and 4 permanent staff for tot. 15 PM</i>
2,7	Equipment	984,97 €	<i>n.3 Personal computer depreciation for the 1st period (total amount of the n.2 invoices € 6.277,62)</i>
2,7	Travel	2.611,45 €	<i>Mission expenses</i>
TOTAL DIRECT COSTS		79659,72 €	

Personnel

Vittorio Murino, Full Professor, person responsible of the work, is charged for 4 PM, Umberto Castellani (researcher) for 3 PM, Manuele Bicego (researcher) for 5 PM and Marco Cristani (researcher) for 3 PM.

Dong Seon Cheng, post-doc, since 1st January 2009, totally paid with SIMBAD funds (3 PM).

Cerruti Stefania, PhD Student, since 10th November 2008, totally paid with SIMBAD funds (4 PM).

Travels

Cristani Marco, Vittorio Murino, Marcella Bellani: Participation to the kick-off meeting in Venice (April 2008)

Bicego Manuele, Castellani Umberto, Cheng Dong Seon: Participation to the Project Meeting in York (November 2008)

Murino Vittorio, Castellani Umberto, Cheng Dong Seon: Meeting with the ETH Partner for project related research activities in Zurich (March 2009)

Report on additional resources

Anna Carli, PhD student, not paid with SIMBAD funds but involved in the research activity, 2 PM

Alessandro Perina, PhD student, not paid with SIMBAD funds but involved in the research activity 3 PM

Marcella Bellani, PhD student, not paid with SIMBAD funds but involved in the research activity 2 PM

TABLE 3.1 PERSONNEL, SUBCONTRACTING AND OTHER MAJOR DIRECT COST ITEMS FOR BENEFICIARY 6, EIDGENOESSISCHE TECHNISCHE HOCHSCHULE ZUERICH, FOR THE PERIOD FROM 01/04/2008 TO 31/03/2009			
Work Package	Item description	Amount	Explanations
3, 6	Personnel costs	35.156 €*	<i>Salaries of 2 doctoral students (first 4.5 months, second 4 months)</i>
3,6	Other Direct Costs	3.065 €*	<i>Travel expenses (Venice, York, Heidelberg, Basel)</i>
TOTAL DIRECT COSTS		38222 €*	

Personnel:

Sharon Wulf, 4,5 PM: since Nov. 15, 2008 PhD student

Peter Schüffler, 4 PM: since Dec. 1, 2008, PhD student

Travels:

Joachim Buhmann and Thomas Fuchs participated in the SIMBAD-meetings in Venice and York
 Thomas Fuchs gave a presentation about SIMBAD-results: "Machine Learning approach to tissue micro-array (TMA) analysis" at the university in Heidelberg.

Sharon Wulf and Peter Schueffler took part in informal meetings at the University Basel / Group Volker Roth.

Report on additional resources:

The following people are not paid with SIMBAD-funds, but they are involved in the project:

Joachim Buhmann, full professor, head of ETH group; Volker Roth, associate professor at University Basel; Dr. Cheng Soon Ong, research associate; Thomas Fuchs, PhD