FET Open short proposal

A new paradigm for large scale manifold learning

LSML

**Funding scheme:** Small or medium-scale focused research project (*short proposal*)

**Workprogramme topics addressed:**
Section 1: Scientific and/technical quality, relevant to the topics addressed by the call

1.1 Targeted breakthrough and its relevance towards a long-term vision

Describe the breakthrough(s) that you are targeting to achieve.

This project aims at developing an emerging paradigm shift in the part of machine learning that goes under the name of “manifold learning”. In many domains it has long been observed that data typically reside in a much lower dimensional manifold. Researchers, though, are becoming increasingly aware of the limitations of traditional “adaptive” spectral embeddings, based on computing an affinity matrix from a graph constructed from a training set of unlabelled datapoints. Most real world problems involve a huge number of such data points, which typically live in extremely high-dimensional spaces. In such conditions spectral embeddings are just not computationally feasible.

A radical new, counterintuitive way of understand the structure of the data in our possession consists on coding it in a random way, against any common sense and the traditional way of thinking the problem. Such linear, sparsity-oriented, metric-preserving “non-adaptive” random projections to a much lower-dimensional “measurement” space have the potential of tackling large scale manifold learning, greatly widening its scope and making pervasive computing scenarios possible. However, adaptive embeddings possess highly desirable properties a large scale manifold learning framework should retain.

The admittedly ambitious goal of this project is to exploit the rising tide towards compressed sensing to design a novel paradigm for large scale manifold learning which integrate adaptive and non adaptive methods in a coherent scheme, in which new classification and clustering techniques are developed in the measurement space and deployed to solve problems simply not accessible before.

What is the long-term vision (scientific, technological, societal, other) that motivates this breakthrough?

The complexity of world of today is reflected by the increasing difficulty of managing the sheer of amount of information any decision maker is confronted with. Human policy makers could potentially exploit the quality and quantity of the data available to them, but they are simply overwhelmed by it and risk neglecting crucial information when making important decisions. A sensible principle (applied for instance in semi-automatic surveillance) would be to gather and interpret data automatically, and escalate to human intervention only in exceptional cases. However, increasingly diffused machines such as autonomous sensors are also confronted with the task of putting some order in the data they gather in order to react effectively to new events and situations.

The recent ISTAG report on Europe’s ICT strategy outlines a vision of the future in which “Products with totally new capabilities will become available for general use, dreams such as intelligent cars, … homes sensitive and responsive to the needs of the persons living in it will become available to everybody. … we will be able to be fully mobile and at the same time stay constantly in touch with everybody.”

All such dreams implicitly assume the capability of handling large amount of data and making sense out of it.

Effective interpretation of large scale flows of data is therefore crucial in order to pursue any such ambitious plans for the future of our society and the way we are part of it.

In “A Preparatory Workshop for EU Seventh Research Framework Programme” held in Luxembourg on December 2005 on Machine Learning and Cognition, the participants “strongly felt that mathematical foundations are important in this area of research” and pointed out that “research could also address the use of optimization methods to assist in the analyses of new representations […] for reducing the complexity of data in high-dimensional spaces or over time”. This is precisely the aim of this project.

Its implementation, the dissemination of its results in both academia and the wider societal context will likely set new trends in the way we discover and represent the structure of data we observe, with consequences on a number of scientific, policy and societal issues it is hard to overestimate.

The scientific community is at this very moment starting to move in this direction. The very purpose of the FET-Open initiative, according to the recent “ICT Research – New perspectives” commission report is to “react to new ideas and opportunities as they arise from within the scientific community or society”. This is the right moment to seize, as the community’s awareness of the need for new, radical ways of learning from and making decisions based large scale data is arising. The Lisbon Strategy for growth and jobs explicitly mentions the need for Europe to keep open to the most recent developments in ICT.
Failure to lead the mounting tide in large scale data management and learning, an issue at the root of most components of the connected society of the near future would compromise the continent’s competitiveness and put serious obstacles before the realization of such scenarios.

Explain how this breakthrough is an essential step towards the achievement of your long-term vision, in particular in terms of new forms and uses of information and information technologies.

The long-term vision we just exposed is consistent with a scenario of societal computing (see ISTAG report), in which all facets of society are managed through ICT in order to optimize outcomes of both societal and economic nature. The novel approach to learning the inherent structure of large scale data we propose here is the key to the actual implementation of many if not all the ramifications of this scenario. This project will fill a gap in the main ICT research streams, which put much emphasis on new technologies, and less on the foundational paradigm behind these technologies, as FETs are indeed supposed to do. In fact, the proposed research also complements the list of priorities given by the FET-Proactive initiative.

The internet of the future will be as much about connecting things as about connecting people. Autonomous, self-organizing systems (such as, for instance, self-organizing traffic lights) will adapt autonomously to changing requirements, reducing the need for human intervention or even centrally-directed autonomous planning. Compressed, non-adaptive sensing could be crucial in reducing (without losing information) the amount of data to store and transmit between different agents/sensors, making these scenarios realistic.

Among the declared objectives of the commission is indeed the exploitation of existing EU strengths: embedded systems are one of them. The potential of non-traditional manifold learning techniques in allowing efficient communication between such systems is apparent. The application of efficient coding schemes has clear consequences on the sustainability of this future network of intelligent agents. The usage of scarce resources such as storage and time is minimized by compressed sensing and its integration in a novel hybrid learning approach, contributing to reduce our carbon footprint in ways that will be proportional to the future actual realization of these scenarios.

A paramount application of such networks of sensors concerns the natural interaction (including surveillance, assistance, etc.) between humans and machines. One of the bottlenecks in the use of ICT occurs when people and computers have to communicate. Streamlining such interactions will have enormous impact on productivity and as a consequence on economic growth. Images and videos from visual sensors provide a wealth of information with the potential of allowing much more natural forms of interaction. This data, however, is bulky and expensive to store or transmit: its efficient but faithful representation could allow applications currently forbidden because of the sheer scale of the data involved.

Support to decision making and customization in business is another important area in which intelligent management of large amounts of data can have an impact. Using real time detailed data, products can be customized for individual clients, allowing companies to offer tailored services and boosting their competitiveness. Efficient methods of organizing such large scale data, and making inference on them are then required.

Describe the concrete objectives that you would consider to constitute the proof-of-concept of such a breakthrough.

With this scenario in mind, the concrete objectives we propose to pursue are the following:

- Study the causes of computational limitations in traditional adaptive manifold learning approaches, and the integration of compressed sensing with traditional adaptive embeddings in a hybrid manifold learning scheme.
- Explore the metric and geometrical properties of hybrid techniques, in relation to the computation of invariants for shape recognition, or unsupervised clustering techniques.
- Develop, to moderate risk, alternative solutions to the large scale manifold learning problem, by studying both approximate spectral decomposition algorithms and approximate nearest neighbor methods to build the affinity graph of the adaptive component.
- Exploit the general-purpose hybrid manifold learning framework developed within the project to propose new solutions to a number of cornerstone problems of machine learning, such as classification, matching, clustering.
• Apply the array of techniques developed during the project to challenging real-world applications in computer vision (object recognition), human motion analysis, medical imaging.

The objectives should be those that you consider achievable within the project, in spite of the inherent risks.

Some of elements involved in the objectives listed above involve indeed high risk: compressed sensing is a discipline still in its infancy, and its proposed integration with adaptive manifold learning is an ambitious task we set for ourselves. Nevertheless, the competences and skills of our consortium give us the confidence to accept such a challenge. The principled formulation of classification and clustering schemes in the reduced, measurement space is open territory, with only few sporadic papers appearing on these topics at the moment. Other issues have recently studied by some groups for traditional embeddings: this is the case for the geometry of the associated eigenfunctions and their invariance. Members of the consortium do possess such expertise. However, such matters are far from being fully understood even in the case of adaptive manifold learning. Actions to moderate risk have been included, based on more traditional approaches to large scale manifold learning. Finally, a number of applications that fall in the wide spectrum of expertise present in the consortium ensures a balanced mix of activities between high, medium, and low risk enterprises.

1.2 Novelty and foundational character

% state of the art, issues of traditional embeddings, the large scale problem, rising discipline of compressed sensing

Even though machine learning problems normally concern data leaving in extremely high-dimensional spaces (e.g., think of natural images in computer vision), it has long been observed that, for a given problem, data typically reside in a much lower dimensional manifold. A number of methodologies have been developed in the last ten years in which such the coordinates on such unknown manifold are computed for a number of sample data, forming a training set. Many of these algorithms are based on eigenvector decomposition of an affinity matrix; such matrix can either measure geodesic distances between all pairs of sample points in a graph approximating the manifold (e.g., ISOMAP) or consider only local reconstruction weights in small neighborhoods (e.g., graph Laplacian, LLE). More recently alternative methods which do not rely on computing such an expensive affinity matrix, but estimate a sparse collection of local, linear bases have been proposed. In both cases training set can be partially labeled, allowing classification directly in the lower dimensional manifold.

Training set, adaptive based embeddings have many desirable properties. Among these, the link between such spectral embeddings and the geometry of the original cloud of data points has been attracting considerable interest. The embedded cloud is invariant to interesting classes of deformations, such as articulated motion, and can therefore be used to compute object signatures to classify the object represented by the cloud. The nodal sets of eigenfunctions of the affinity matrix are in close relation with symmetries and other topological properties of the underlying manifold.

Adaptive embeddings, though, suffer of major computational problems in terms of both the dimension of the original data space, and the number of samples. Traditional methods settle for a tradeoff between number of samples and computational cost, even though Hebbian methods for incremental SVD have been proposed to tackle this issue. The complexity in the number of points can be reduced by using locally isometric methods, but these do not preserve all distances anymore, so that, for instance, kernels in the embedding space are not consistent with kernels in the original space.

What is the novelty of your proposal? In what way do you challenge current thinking or assumptions? Novelty should come from new ideas, not from the incremental refinement of existing approaches. It can also come from new and unexpected combinations of insights from various disciplines.

Very recently, non-adaptive approaches to manifold learning have been proposed, as people discovered that important metric properties of an underlying manifold could be preserved, with high probability, by a sufficient number of linear orthonormal projections sampled from simple Gaussian (amongst others) distributions. Such mappings: do not depend on the choice of a particular training set; are very fast, as they are linear and randomly generated, not involving expensive graph constructions or eigendecompositions; preserve not only distances, but also scalar products and angles; their complexity
depends linearly on the dimensionality of the reduced space $K$, and only logarithmically on the dimensionality of the original space $N$. The community is still at an early stage in recognizing the good properties of random projections, and how they could be exploited to develop radically new manifold learning frameworks, in both compression, classification, and clustering.

The idea that mappings do not need to be training-set driven in order to preserve fundamental properties of the data is a radical challenge to the current way of thinking the problem, and is likely to develop into an emerging field with the potential of become the new mainstream in manifold learning.

**What is the scientific foundation that you aim to develop?**

In this project we propose to employ and further develop the rising field of random projections in compress sensing to overcome the distinctions between adaptive and non-adaptive algorithms for unsupervised manifold learning. Our aim is to develop a radically new hybrid framework, in which traditional training set based embeddings are integrated with random projection to deliver an effective solution to the large scale manifold learning problem.

A principled fusion of the two approaches would: eliminate barriers to the immense potential applications of manifold learning due to computational complexity; preserve a host of nice geometric properties, such as transformation-invariance and symmetries, exhibited by graph embeddings. Such a framework would open the door to new, exciting applications. As random projections preserve not only distances but also scalar products, the application of support vector machine like techniques to perform classification directly in the reduced space become possible.

**1.3 S/T methodology**

Provide an outline of the scientific and technological approach or methodology by which you will attempt to reach your objectives.

The project will articulate into a three-stage structure, designed to fit the ambitious goals of studying and developing a novel paradigm for large scale manifold learning, and propose off-the-shelf machine learning tools to a wide audience of scientists, engineers, and practitioners. The first area concerns the theoretical development and study of a coherent hybrid manifold learning framework, building on the existing properties of adaptive and non-adaptive manifold learning techniques. In particular, the crucial geometric properties of the resulting hybrid schemes will be investigated to explore their full application potential. To moderate risk, we will pursue in parallel more traditional approaches to large scale manifold learning such as approximate spectral decompositions and approximate graph construction prior to embedding.

In the second (logical, but not temporal) stage of the project novel machine learning methodologies based on hybrid large scale manifold learning will be developed. Classification, clustering, multilinear modeling, among others, are problems which suffer from the curse of dimensionality, and see their application constrained by running time or data storage limitations. Their formulation in a measurement space after hybrid embedding is an entirely new topic of research which is likely to become mainstream in the near future.

To prove the efficacy of the new developed tools in solving hard, open problems in manifold disciplines which just cannot be tackled properly by means of traditional methods, we will apply such tools to a range of such problems in fields as diverse as object recognition, internal organ simulation, and human motion analysis.

In more detail the project will be articulated into the following research strands, which correspond to our stated objectives.

**Study of compressed sensing and its integration with adaptive embeddings [Brookes?].** The combination of adaptive (based on a number of samples forming a training set) and non-adaptive (in which the computed embedding does not depend on the specific choice of a training set of sample) manifold learning methods is the core of the paradigm shift we propose in this project. A simple cascade of initial random projection and training-set based embedding does guarantee a number of good properties, such as reduced computational cost, preservation of distances, invariance of the transformation to a wide class of deformations of the data cloud of samples, and could be used as a sensible starting
point. The application of compressed sensing itself to manifold learning is still in its infancy, and lots of aspects of such option are still unclear. Very recent results, for instance, seem to prove that the GP (Grassberger-Procaccia) estimate of the inherent, intrinsic dimension of the manifold from which the data are sampled is also preserved by an appropriate number of random projections. The residual variance of applying, say, ISOMAP in the measurement space can also be forced to be close to the same quantity in the original space (by setting a sufficient number of orthonormal projections). This ensures that the “quality” of the learnt manifold as a good interpolation of the given training set is the same after and before random projection.

In addition, very recent even though preliminary results seem to indicate that not only distances, but scalar products and angles are also preserved in the reduced space, allowing classification in this reduced space through, for instance, SVM. It is worth noticing that semidefinite embedding (SDE) does the same, but at a prohibitive cost. A full understanding of the geometrical features of hybrid manifold learning scheme will be a challenging but exciting task.

An entirely open question is what to do when data points are actually vectors formed by non-homogeneous entries coming from entirely different sources (say, appearance and shape in object tracking). Distance functions which equally weigh different entries of the vectors (such as the standard Lp distances) do not make sense in this case. This issue is crucial whenever the structure of data after information fusion is to be analyzed. We will study it by building on recent interesting results on so called “joint manifolds”.

Geometrical properties of the proposed hybrid frameworks. [INRIA/Technion] A striking feature of adaptive embeddings is that they inherent important geometric properties of the training cloud of data points, as a finite approximation of the underlying, unknown manifold. This is a consequence of spectral decomposition via SVD. Preliminary studies, conducted mainly within the graphics community, indicate that the eigenvalues of the affinity matrices constructed via graph Laplacian, for instance, are functions of the topological invariants of the manifold. Moreover, the embedding cloud is invariant to interesting classes of deformations, such as rigid and articulated motion. This allows to solve the data association problem directly in the embedding space by simple orthogonal alignment. In the last few years researchers have also noticed that the zero-level set or “nodal sets” of the eigenvectors of the affinity matrix are located in correspondence of the main symmetries axes of the original cloud of data points. Despite the growing interest around graph-based embeddings, a full explanation of their attractive geometric feature has proved elusive. However, as random projections preserve (with high probability) pairwise distances after projection in a much lower dimensional measurement space, a (for instance) cascade combination of random compression and adaptive embedding will likely preserve such geometrical features. The exact characterization of these conservation results will be a cornerstone of this strand of our research.

Alternative solutions to the large scale manifold learning problem [Tubingen?]. As we are acutely aware of the foundational nature of the studies, and the associated risk, we believe it will be wise to pursue in parallel alternative approaches to the issue of large scale manifold learning. Very recently approximate spectral decomposition algorithms have been proposed to reduce the computational complexity of the SVD-like stage common to many embeddings schemes, such as the Nystrom method or column sampling approximation. Even though such methods are based on random selection of columns of the affinity matrix, they are very different in nature from the random projection paradigm. Properties and limitations of such approximate schemes are not yet understood. Any knowledge gained from their study could prove crucial in the development of a coherent hybrid framework.

In addition, it has been noticed that the other serious bottleneck of traditional embeddings is the computation of a graph approximating the given training set. This in turn requires the detection of a (usually fixed) number of nearest neighbors for each sample point. Fast nearest neighbor algorithms such as PSH could be the key to dramatically improving the performance of adaptive embeddings in large scale contexts. Open questions concern the desirable properties of adaptive embeddings after such approximations: which of these features survive?

Development of machine learning algorithms after hybrid embedding [Brookes/Technion/INRIA?]. The crucial properties that a hybrid scheme inherits from that of non-adaptive random projections and adaptive graph-based embeddings open enormous possibilities of radically influencing numerous fields of machine learning such as classification, shape analysis, and clustering.

Classification. SVMs are very effective and popular tools for supervised classification. However, their
applicability is seriously limited by the huge number of features involved, for instance, in challenging vision problems such as scene understanding. Very recently the possibility of designing support vector machines directly in the measurement space after appropriate random projections has emerged. This is due to the Restricted Isometry Property (RIP), which ensures that pairwise distances are preserved in the measurement space with high probability. This has potentially groundbreaking consequences for the entire field. 

**Invariant shape signatures.** As geometric features of adaptive embeddings are inherited by hybrid schemes, fast invariants can be computed in the measurement space for classification purposes. As the embedded cloud is invariant to important classes of transformations, shape invariants can be computed there in order to classify non-rigid shapes in a robust and efficient way. Similar approaches such as shape DNA have been attempted using traditional embeddings. The community’s understanding of exactly which classes of deformations guarantee invariance is still embryonal. 

**Clustering.**

**Applications of developed algorithms to computer vision, motion analysis, medical imaging [INRIA, Brookes, UPF].** We plan to test the developed hybrid learning framework and the accompanying array of novel machine learning tools to a wide, interdisciplinary range of challenging real world problems, exploiting the range of expertise available within the consortium. 

**Human motion tracking and recognition** is one such problem, in which the complex structure of the moving body, the presence of other objects in the scene, and self occlusions make tracking the position of the moving person extremely difficult. The invariance of hybrid embedding under articulated motion can then be used to match the points belonging to the human along a video sequence, based on their shape and appearance. 

In **scene understanding**, the goal is to classify each pixel in an image, or a sequence of images, to a certain class of objects. Perhaps the most critical obstacle in this context is the massive number of image observations (usually called “features”) and variables to estimate when, for instance, you want to attribute a certain label (object class) to each pixel in the available images or an entire video sequence. This forces to use underperforming optimization methods such as boosting to solve the problem instead of state of the art SVM approaches. Dimensionality reduction with adaptive embeddings is impractical when observations live in a space of dimension of an order of magnitude around 10-100 thousand, and data points number around 100 million. Compressed sensing can cope with the problem by drastically reducing in a non-adaptive way the dimensionality of the data, while preserving pairwise distances. This opens the way to a direct application of the SVM framework instead of boosting, with a considerable improvement in performance. 

In **medical imaging and simulation**, internal organs and blood vessels are to be simulated, starting from the available data in the form of layered MRI scans. Normally such organs (such as the heart) are represented by clouds of samples points by point distribution models, whose vertices are then tracked according to image data along a sequence. The use of non-linear manifolds for constraining the plausibility of shapes represented by point distribution models has not been explored so far. Exceptions are some early, very specific attempts to kernel PCA or mixture of Gaussians. A practical implementation of this idea involves serious issues, especially when dealing with large datasets. As traditional manifold learning methods rely on training data, to determine whether a new PDM is a valid instance of a certain object's shape you need to go through an expensive procedure of finding the neighbors and, perhaps, locally approximate the manifold. The real challenge is to find a basis that allows to compute a mapping function to- and from- the manifold, so that we can discard the training data. Methods to learn sparse local bases which approximate a non-linear manifold are beginning to appear. Such approaches would be easily tested against typical PDM constraints in the 3D organ tracking problem. 

*Where relevant, show how your approach is well-adapted to the multi-disciplinary nature of the novel idea or approach that you are proposing.*

As it is apparent from the organization of the project into a number of interacting but autonomous strands, we aim not only at defining a radically new paradigm for large scale manifold learning, but also at developing the intrinsic potential of this framework. Foundations, machine learning techniques and applications are all tackled in a coordinated effort to organize this emerging field into a coherent whole. Periodic coordination actions will be taken to ensure coherence is maintained throughout the project’s lifespan.
For all of the above applications the consortium is already in possession of significant data sets gathered throughout the last years. Additional, more targeted sets of experimental data will be acquired in the course of the project, thanks to the availability of motion capture systems, collections of multiple synchronized cameras for 3D reconstruction, and the collaboration with hospitals. All such data sets will be annotated with ground truth data, and made available over the internet together with the generated code to maximize the impact of such an ambitious project.

*Provide evidence that you are aware of the level and nature of the risks of failure, and that you have a good idea on how to address these risks.*

As we mentioned above, the project is articulated into different layers, each coming with a different level of risk. While a coherent formulation of hybrid manifold learning combining the desirable features of adaptive and non adaptive algorithm is a challenge that cannot be met by simply building on existing results, activities in the second layer are characterized by a lower level of risk, while applications will surely produce knowledge in terms of powerful insights on the problem. Overall, the project is designed as a pinch in which several arms attack the problem from different angles. Risk is further moderated by pursuing in parallel alternative approaches to the large scale manifold learning problem.

*What would constitute success? What would you learn from an eventual failure?*

Success will be measured by quantifying the performance of the classification, clustering and estimation algorithms devised in this novel framework on the large scale data sets in possession of the consortium. Performances will also be measured on a wealth of publicly available data sets in the fields of object recognition (e.g. the PASCAL challenge), video segmentation and understanding, medical data analysis (CITE), for which ground truth is available and well-established methodologies for assessing effectiveness of algorithms exist. It is important to notice that the focus on large scale approaches will enable us to tackle applications that were just not possible to cope with before. A direct consequence of these results will be the publication of our research on prestigious international journals and the major international conferences in Machine Learning, Computer Vision, and Medical Imaging. A direct measure of our success in developing hybrid manifold learning into a coherent field will be the publication of the core results on general interest journals such as Science or Nature.