Small or medium-scale focused research project (STREP)

ICT Call 9
FP7-ICT-2011-9

[Dynamical Generative and Discriminative Models for Action and Activity Localization and Recognition]

Type of project: Small or medium scale focused research project (STREP)
Date of preparation: 17/04/2012
Version number (optional):

Work programme objective addressed: Objective 2.1 - Cognitive systems and robotics
(if more than one, indicate their order of importance to the project. The main (first) objective must be one included in this call)

Name of the coordinating person: [Dr Fabio Cuzzolin]
e-mail: [Fabio.Cuzzolin@brookes.ac.uk]
fax: [+44 (0)1865 484545]

<table>
<thead>
<tr>
<th>Participant no.</th>
<th>Participant organisation name</th>
<th>Part. short name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Coordinator)</td>
<td>Oxford Brookes University</td>
<td>OBU</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>2</td>
<td>Dalle Molle Institute for Artificial Intelligence</td>
<td>IDSIA</td>
<td>Switzerland</td>
</tr>
<tr>
<td>3</td>
<td>Ghent University, SYSTeMS Research Group</td>
<td>SYSTeMS</td>
<td>Belgium</td>
</tr>
<tr>
<td>4</td>
<td>Ecole Supérieur d'Electricité</td>
<td>SUP</td>
<td>France</td>
</tr>
<tr>
<td>5</td>
<td>Dynamixyz</td>
<td>DYN</td>
<td>France</td>
</tr>
</tbody>
</table>

Proposal abstract

Action and activity recognition are intuitive but extremely difficult tasks, which lie at the root of a panoply of scenarios of human-machine interaction, ranging from gaming, mobile computing and video retrieval to health monitoring, surveillance, robotics and biometrics. The problem is made challenging by the inherent variability of motions carrying the same meaning, which in turns causes over-fitting issues due to the forcibly limited size of any available training set, but also by the presence of numerous nuisance factors such as locality, viewpoint, illumination, occlusions, and many more. While recent bag-of-features-like approaches have focused mainly on classifying histograms of features extracted from spatio-temporal volumes (completely ignoring, in this way, the causal structure of motions to recognize), dynamics can provide invaluable information in order to successfully locate and recognize actions and complex activities in a robust and reliable way. Traditional generative dynamical models, however, have proved unable to cope effectively enough with some of these difficulties.

We propose here to design and test novel frameworks for the integration of action dynamics in both generative and discriminative models, with the aim of stimulating a breakthrough in activity recognition, with enormous exploitation potential in the manifold scenarios indicated above. Novel classes of generative graphical models based on the principle of allowing the probabilities defining the model to vary within whole convex sets rather than assuming sharp, precise values are formulated in order to address the issue of overfitting due to the limited size of the training sets. New manifold learning techniques are applied to generative graphical models in order to tackle the presence of numerous nuisance factors. Finally, a new class of discriminative, part-based models originally developed for object recognition are generalized to action localization and recognition as a fundamental way of coping with complex activities formed by series of simple actions, and addressing the issues of locality and presence of multiple actors.
Table of contents

1 Scientific and/or technical quality, relevant to the topics addressed by the call 3

1.1 Concept and objectives 3
   1.1.1 Topic of research 3
   1.1.2 Research hypothesis and ideas 4
   1.1.3 Scientific objectives 6
   1.1.4 Relation to the topics addressed by the call 6
   1.1.5 Feasibility and timing 7

1.2 Progress beyond the state-of-the-art 7
   1.2.1 Current challenges in action recognition 8
   1.2.2 Advantages and disadvantages of current state-of-the-art approaches 8
   1.2.3 Contributions: Pushing the boundaries in generative and discriminative approaches 10

1.3 S/T methodology and associated work plan 11
   1.3.1 Breakthroughs 11
   1.3.2 Overview of the S/T Methodology 12
   1.3.3 Detailed description of the Work Packages 13
   1.3.4 Milestones 21

2 Implementation 37

2.1 Management structure and procedures 37
   2.1.1 Organisation 37
   2.1.2 Communication strategy 40

2.2 Individual participants 41
   2.2.1 – OBU : Oxford Brookes University 41
   2.2.2 – IDSIA : Istituto Dalle Molle di Studi sull'Intelligenza Artificiale 42
   2.2.3 – SYSTeMS : Control systems group at Universiteit Gent 43
   2.2.4 – SUPÉLEC : Ecole Supérieur d'Electricité 44
   2.2.5 – DYNAMIXYZ 45

2.3 Consortium as a whole 46
   2.3.1 Past relevant results by the partners 46
   2.3.2 How the participants constitute a consortium capable of achieving the objectives 47
   2.3.3 Suitability and commitment to the tasks assigned 48
   2.3.4 Complementarity between participants and overall balance 48

2.4 Resources to be committed 49

3 Impact 51

3.1 Expected impacts listed in the work programme 51
   3.1.1 Connection with expected impacts in the work programme 51
   3.1.2 External factors affecting the potential impacts 53
   3.1.3 Why a European approach 54

3.2 Dissemination and exploitation, management and intellectual property 55
   3.2.1 Beneficiaries of the proposed project 55
   3.2.2 Measures proposed for the dissemination/exploitation of the results 56
   3.2.3 Timescales 60
   3.2.4 Intellectual Property management 60

4 Ethical Issues 61

Annex I: References 63
Section 1: Scientific and/or technical quality, relevant to the topics addressed by the call

1.1 Concept and objectives

1.1.1 Topic of research

Action and activity recognition. Since Johanssen's classic experiment showing that moving light displays were enough for people to recognise motions or even identities [110], recognising human activities from videos has been an increasingly important topic of computer vision. The problem consists in telling, given one or more image sequences capturing one or more people performing various activities, what categories (among those previously learned) these activities belong to. A significant (though semantically rather vague) distinction is that between "actions", meant as simple (usually stationary) motion patterns, and "activities" considered to be more complex sequences of actions, sometimes overlapping in time as they are performed by different parts of the body or different agents in the scene.

Why is action recognition important: scenarios for real-world deployment. The potential impact of a real-world deployment of reliable automatic action recognition techniques is enormous, and involves a large number of scenarios in different societal, scientific and commercial contexts (Figure 1). Historically, ergonomic human-machine interfaces able to replace keyboard and mouse with gesture recognition as the preferred means of interacting with computers may have been envisaged first. It has been stated that 93% of human communication is non-verbal, that is, with body language, facial expressions, and tone of the voice imparting most of the information involved. In this scenario humans are allowed to interact with virtual characters for entertainment, educational purposes, or at automated information points in public spaces. Smart rooms have also been imagined, where people are assisted in their everyday activities by distributed intelligence in their own homes (switching lights when they move through the rooms, interpreting their gestures to replace remote controls and switches, etcetera). In particular, given our rapidly ageing population (at least in Europe), semi-automatic assistance to non-autonomous elderly people (able to signal the need for intervention to hospitals or relatives) is becoming an object of increasing interest.

In the surveillance scenario, in contrast, the focus is more on detecting “anomalous” behavior in public spaces, rather than precisely determining the exact nature of the activity that is taking place. Rather than spending a lot of time (and money) in front of an array of monitors, security guards can be assisted by semi-automatic event detection algorithms able to signal anomalous events to their attention. In the related field of biometrics, popular techniques such as face, iris, or fingerprint recognition cannot be used at a distance, and require user cooperation. For these reasons, identity recognition from gait is being studied as a novel “behavioral” biometric, based on people’s distinctive gait pattern.

The newest generation of controllers (such as the movement-sensing Kinect console by Microsoft) have opened new directions in the gaming industry. Yet, these tools are only designed to track the user’s movements, without any real elaboration or interpretation of their actions. Action recognition can "spice" up the console owners’ gaming experience to their great satisfaction.

Last but not least, the availability of enormous quantities of images and videos over the internet is one of the major features of nowadays' society. People acquire new videos all the time with their own smartphones and portable cameras, and post them on Facebook or YouTube. However, techniques able to efficiently data-mine them are a scarce resource: the potential of a "drag and drop" style application, similar to that set up by Google for images, able to retrieve videos with a same "semantic" content is easy to imagine.

Figure 1. Action and activity recognition have manifold applications: virtual reality, human-computer interaction, surveillance, gaming and entertainment are just some examples.

Crucial issues with action recognition. While the formulation of the problem is simple and intuitive, action recognition is, in fact, a hard problem (Figure 2). Motions possess an extremely high degree of inherently variability: quite distinct movements can carry the same meaning or represent the same gesture. In addition, they are subject to a large number of nuisance or “covariate” factors [37], such as illumination, moving background, camera(s) viewpoint, and many others. This explains why experiments have been often
conducted in small, controlled environments, even though attempts have been recently made to go towards action recognition “in the wild” [37]. If we remove the assumption that a single motion of interest is present in our video(s), locality emerges too as a critical factor: a person can walk and wave at the same time; different actors may perform different activities in the same room. Indeed, action localization (in both space and time) within a given video sequence is a necessary first step in any recognition framework. However, most current research normally assumes that the motion of interest has already been segmented, and focuses on the classification of the given segment. This is reflected in the inadequate number of action localization/detection (versus recognition) test-beds currently available.

Current research tends as well to focus on single-actor videos, as the presence of multiple actors greatly complicates both localization and recognition. In addition, learning a large, realistic number of action categories requires the collection of huge training sets, and triggers the problem of how to classify large amounts of complex data. Most importantly, learning action models from a limited training set poses serious overfitting issues: if very few instances of each action category are available the model describes well the available examples, but has limited generalisation power, i.e., it performs badly on new data.

Finally, a serious challenge arises when we move from simple, “atomic” actions to more complex, sophisticated “activities”, seen as series of elementary actions structured in a meaningful way.

Figure 2. Issues with action recognition, from left to right: viewpoint, trajectory, locality and context, multiple actors and spatio-temporal localization.

1.1.2 Research hypothesis and ideas

The role of dynamics in localization and recognition in more realistic settings. Recently, recognition methods which neglect action dynamics (typically by building bag-of-features (BoF) [6] models extracted from the spatio-temporal volume associated with a video [32]) have proven rather effective (http://www.wisdom.weizmann.ac.il/vision/SpaceTimeActions.html). However, those techniques have been typically tested on datasets composed by pre-segmented action instances, depicting rather elementary actions, and contemplating only a rather limited number of action categories. To move beyond the current setting, and bring action and especially activity recognition closer to actual deployment in the manifold real world scenarios illustrated above, empirical evidence suggests that modelling dynamics is crucial.

First of all, ignoring the spatial and temporal structure of the motion of interest makes localizing the action embedded in a larger video sequence very difficult. Some insight may come from the object detection problem in 2D images: there, part-based discriminative models in which objects are seen as constellations of parts have proved very successful in localizing them in context. Besides, a simple mental experiment shows that bag-of-features models fail to discriminate videos containing real meaningful actions from scrambled versions obtained by permuting the ordering of the frames!

Explicitly modelling the structure of complex series of actions is essential to move from recognition of elementary actions to that of more sophisticated activities. Finally, as we are going to see below, modelling dynamics in a sufficiently flexible way allows us to some extent to address the all-important issue of overfitting due to necessarily limited training sets (in both the number of training sequences and the number of action categories there contemplated).

Generative dynamical models, systems of differential or discrete-time equations designed to model a given phenomenon, both deterministically or statistically, have a number of interesting characteristics that can be exploited to address several of the issues with action recognition. In primum, they provide ways of temporally segmenting in an automatic way an action embedded in a larger video sequence, for instance by assessing the likelihood of the sequence been generated by the model at each time instant. Max-margin conditional random fields [56] are able to address locality by recognizing actions as constellations of local motion patterns. In situations where a significant number of people ambulate in the field of view, a paradigm shift from single objects/bodies tracking to approaches in which the monitored crowd is considered as some sort of fluid becomes necessary: dynamical models are well equipped to deal with such a scenario. Moreover, sophisticated graphical models can be exploited to learn in a bottom-up
fashion the “plot” of a footage [28,38], which can be used to retrieve the video over the internet. When encoding actions as generative dynamical models of a certain class, action/activity recognition reduces to their classification. Various classes of models have been proposed to this purpose in the past, ranging from hidden Markov models to linear, nonlinear, stochastic or even chaotic dynamical systems. Unfortunately, classical dynamical models have been sometimes found too rigid to describe complex activities, or prone to overfitting the existing training examples. One reason is that they typically need to estimate single, “precise” probability distributions, e.g. the state transitions of a traditional or a hierarchical hidden Markov model (HMM) or again the distribution associated with each HMM state in the space of observations (see Figure 3).

Figure 3. A classical hidden Markov model (HMM) is defined by a transition matrix \( A \) between its finite states (representing e.g. "canonical action poses") and, for each state \( j \), the conditional probability \( p(y|j) \) of generating an observation \( y \) (image/video feature) given state \( j \). In response to the uncertainty induced by having to learn a model from a training set of limited size, its imprecise-probabilistic version (iHMM) allows entire convex sets of transition and emission probabilities rather than single probability distributions.

Modelling dynamics via imprecise-probabilistic graphical models. How to retain the desirable features of generative dynamical models, while addressing their troubles with overfitting and rigidity? The theory of imprecise probabilities [82] and the extension of graphical models to allow the manipulation of whole convex sets of probabilities [81] provides a way of overcoming overfitting issues, while allowing the resulting classifier to handle missing or partial data in a natural way. Indeed, in decision making and estimation the state of the world is normally assumed to be described by a probability distribution over a set of alternative hypotheses, which we are required to estimate from the available data. Sometimes, however, as in the case of extremely rare events (e.g., a volcanic eruption), few statistics are available to drive the estimation. Part of the data can be missing (think of occlusions). Furthermore, under the law of large numbers, probability distributions are the outcome of an infinite process of evidence accumulation, while in all practical cases the available evidence only provides some sort of constraints on the probabilities governing the process. This is precisely the situation in action recognition, in which generative or discriminative models have in practice to be built from just a small fraction of the evidence theoretically at our disposal (think of all the videos stored on the internet). Different constraints are associated with different generalisations of probabilities, such as interval probabilities [80], convex sets of probabilities or credal sets [81], Shafer's theory of belief functions [83,84], coherent lower previsions [86], and others. The development of a coherent theory of imprecise-probabilistic graphical models allowing the manipulation of whole convex sets of probabilities rather than single distributions has the potential to make generative models more flexible, while retaining their desirable properties.

Modelling dynamics via discriminative models. On the other hand, an action's spatio-temporal structure can be taken into account in the form of discriminative models. The insight here comes from the object detection problem, in which articulated objects composed of different rigid parts (such as human bodies) have been recently modelled as constellations of "parts" which are learned in a discriminative way, achieving state-of-the-art results [115,116]. This hierarchical and compositional nature of human actions, instead, has not yet been exploited by current BoF approaches. Some work in this sense has been done [114]: BoF models can be used to represent action-parts, since
build artificial systems that can reason on what they see and interpret the corresponding, evolving scenes in
objectives: its overall goal is to push the boundaries in learning and recognizing patterns in video data; to
All these crucial elements and issues are contemplated in the propose project, and form indeed its core
important goal and feature of the present project, in fully agreement with the Call. In particular:
Enabling cognitive capabilities in artificial (autonomous or semi-autonomous) systems is
data, allowing the development of autonomous systems in the scenarios we described.

The goal of this project is to develop an all-purpose, multi-modality robust action recognition framework
and robotics"
time; to allow them to react appropriately when anomalous events take place, either directly or by assisting a human operator. Different strategies for automated learning are our landmark contributions. In addition:

- Developing suitable benchmarks, conducting benchmarking exercises and supporting scenario-based competitions are therefore firmly placed on the agenda.

Testing and validating the developed techniques in a variety of different scenarios, assessing their performance over established benchmark test-beds, and gathering new test-beds when necessary to go beyond the current state of the art is the strategy at the heart of the experimental part of this project.

- .. research will be motivated, guided and validated by realistic, demanding and scalable real-world scenarios, where appropriate backed by industrial stakeholders. Gearing up cross-fertilisation between relevant industry and research communities is a key issue in this respect and industrial participation is therefore greatly encouraged.

Our research is directly motivated by the ambition of breaking new ground in action recognition and pushing towards its real world deployment in the many scenarios outlined above. The challenges we plan to tackle are direct consequences of the issues encountered in these scenarios. Cross-fertilization between research and industry is not only necessary in order to achieve our goals, but also naturally in sight in virtue of the strong industrial partner involved and the consortium’s existing commercial links.

In particular target outcome b), Cognition and control in complex systems, concerns:

- ... technologies based on the acquisition and application of cognitive capabilities (e.g., establishing patterns in sensor data, classification, conceptualisation, reasoning, planning) for enhancing the performance and manageability of complex multi-component and multi-degree-of-freedom artificial systems, also building on synergies between cognitive systems and systems control engineering.

The main goal of this project is indeed pattern recognition and classification, in an extremely challenging problem in which novel approaches are necessary to push towards deployment.

- Realistic, highly demanding, scalable real-world scenarios will motivate and guide research related to targets a) & b), and serve to validate its results.

Real-world scenarios are our anchors and references throughout the lifespan of the project (see WP5).

- Specific Targeted Research Projects (STREP) are particularly suited to high-risk endeavours, breaking new grounds, with high potential rewards.

Breaking new grounds in both theoretical (the development of novel learning techniques suitable to efficiently describe spatio-temporal structures in both generative and discriminative modelling) and applicative terms (a push towards the real-world deployment of action recognition) is the goal of this project.

1.1.5 Feasibility and timing

These objectives can only be achieved by partners possessing the necessary, broad range of skills. These include: expertise in cutting edge theoretical work on imprecise probabilities and graphical models, a strong background in computer vision with a focus on feature selection and extraction, action and activity recognition and experience in their software implementation.

The consortium features partners (IDSIA and SYSTeMS, in particular) who are among the pioneers of the study of probabilistic inference via imprecise Markov and graphical models in general [62-64]. The aim of bringing to full development learning and inference algorithms such as Expectation-Maximisation (EM) [67] for imprecise models is very ambitious, but within reach for two of the top groups in the area. The timescale and resources we request are what we believe are adequate to reach such goals.

The consortium (via OBU) also possesses world class expertise in feature selection and extraction, fast classification of large amounts of images and video via graph cut methods and SVMs [85], action recognition and identity recognition from gait [11,14]. OBU has a wealth of industrial partnerships with top companies such as Sony, VICON, and Microsoft Research Europe which will be extremely valuable to maximise impact, and can provide valuable infrastructure in terms of range cameras, motion capture equipment and autonomous robots. Supélec and Dynamixyz are award winning research centers in realistic human animation and facial gesture representation and recognition. For all these reasons, we are confident, the set ourselves ambitious but realistic objectives, achievable within the lifespan of the project.

The project's objectives are articulated into a number of measurable and verifiable milestones in Section 1.3.4: these milestones are summarized in the Table of page 27, and a graphical description of their location within the wider project is given and discussed in page 35.

1.2 Progress beyond the state-of-the-art

Let us now briefly review the state of the art in action recognition. We think best to organise this survey around the current major challenges (Section 1.2.1), in order to be able to measure advantages and
disadvantages of the state-of-the-art approaches versus their ability to cope with those issues (Section 1.2.2),
to eventually illustrate in what way we intend to progress beyond the current state of affairs (Section 1.2.3).

1.2.1 Current challenges in action recognition

Challenge #1: localization in space and time. As we said the first necessary step in any action recognition framework consists on detecting when and where a semantically meaningful action takes place within a video sequence. In the surveillance scenario, for instance, security has to be told when an interesting/peculiar event has started. In a robotics application, an autonomous vehicle may have to segment from the crowd a single person trying to communicate with it. The problem has been tackled by some authors in the literature, yet the general focus is still on the recognition of pre-segmented videos. In [113] a boosted cascade of classifiers is learnt from spatio-temporal features, and searched over an exhaustive set of action hypothesis at a discrete set of plausible spatio-temporal locations and scales. In [112] the task is split into two, by first detecting and tracking humans to determine the action location in space, to later use a space-time descriptor and sliding window classifier to temporally locate two actions (phoning, standing up). In a similar spirit to [113], the work in [114] uses a search strategy in which space-time video regions are matched to manually created action templates. In all these cases, however, a manual labelling of the spatial and temporal extent of the actions in the training set is required.

Challenge #2: multi-agent recognition. A single person can perform multiple actions at the same time, with different parts of their body. Or, distinct actions performed by different agents [112] can take place independently in different spatio-temporal regions of a video. Think of a gaming/entertainment scenario, in which different video game players sit in front of a single console but play cooperatively or as opponents the same game. In response, Gilbert et al [27] copes with this challenge by using very dense spatio-temporal corner features in a hierarchical classification framework, while Reddy et al [45] propose to cope with incremental recognition using feature trees that can handle multiple actions without intensive training. The problem, however, is still very much open.

Challenge #3: actions “in the wild” and influence of covariates. Due to the strong influence of covariate factors in action recognition, experiments have so far typically considered more or less constrained environments (e.g., single agents moving against a uniform background) though a few initial steps have been made to go beyond these. Liu, Luo and Shah, for instance [37], have taken on the challenge of recognising actions “in the wild”, coping with the tremendous nuisance variations of “unconstrained” videos. Hu et al [59] have investigated action detection in cluttered scenes in their SMILE-SVM framework, while Yao and Zhu [58] have proposed to learn from cluttered real-world videos deformable action templates consisting of a set of shape and motion primitives.

Challenge #4: overfitting due to inherent variability and limited training sets. The other major source of complication when classifying actions is, as we said, the inherent variability of actions of a same category: very different spatial trajectories can have to us the same meaning. Experiments are normally run on datasets which have clear limitations in terms of both the number of action classes present, and the number of instances of motions for each given class. The consequence is that models learnt from such datasets tend to overfit the limited training data provided, and have little generalisation power.

Challenge #5: complex activities versus elementary actions. In order to learn and recognize more complex sequences of atomic motions, forming more sophisticated activities, we need to incorporate some sort of description of the activity’s structure. A recent review can be found in [138]. This ability is most relevant to the smart-home scenario, in which people go about their daily lives, composed of a number of complicated sequences of tasks, which their home is supposed to trace and (to some extent) understand.

Challenge #6: beyond pure recognition: video retrieval. In video retrieval over the internet (implemented, say, as a Google images like drag-and-drop web service) the problem is not quite about recognising action categories. Rather, the purpose is to (automatically) compute a descriptor of the video at hand (either in the form of a verbal description, or of a concise signature), and look for video with a similar description/signature. Sophisticated graphical models, for instance, have been used to learn in a bottom-up fashion the temporal structure or plot of a footage, or to describe causal relationships in complex activity patterns [38]. Gupta et al [28] have worked on determining the plot of a video by discovering causal relationships between actions, represented as edges associated with spatio-temporal constraints.

1.2.2 Advantages and disadvantages of current state-of-the-art approaches

How do current state-of-the-art methods address these challenges? Two main directions of research have been explored in the past: “bag-of-features” (BoF) approaches applied to spatio-temporal volumes, and generative (mostly stochastic) modelling of the temporal structure of an action.

Proposal Part B: page [8] of [67]
**Bag-of-features (BoF) on spatio-temporal volumes methods.** These methods, in which dictionaries are built from clustering features extracted from (typically) the spatio-temporal volume associated with a video sequence (Figure 4), have been very successful in recent time. To cite a few, Kim and Cipolla [32] have proposed a spatio-temporal pattern matching approach in which spatio-temporal volumes are seen as tensors, and an extension of canonical correlation analysis to tensor analysis is used to detect actions on a 3D window search. Bregonzio et al [7] have used a global spatio-temporal distribution of interest points, from which they extract and select “holistic” features over multiple temporal scales. Rapantzikos et al [44] have also adopted dense spatio-temporal features detected using saliency measures, in a multi-scale volumetric representation [105]. Yuan et al [60] have described actions, once again, as collections of spatio-temporal invariant features, and proposed a naive Bayes mutual information maximisation method for multi-class recognition.

**Issues with BoF methods.** These methods have delivered good results in recent years, at least on datasets of fairly limited size, with few action categories. They have been tested, however, mostly on classification test-beds with pre-segmented clips, while localization using pure volume descriptors is rather difficult. It is interesting to note that on a new human motion action dataset dataset with 51 action categories (HMDB51 [104]), bag-of-features models [87] achieve classification results of just over 20%, suggesting that these approaches do not fully represent the complexities of human actions. Kuehne et al. has shown that this drop in performance with respect to previous datasets is most likely due to the increased number of action categories. [104] has also highlighted the need for more challenging datasets: while in the UCF Sports dataset [109], human static joint locations alone are sufficient to classify sports actions with an accuracy of over 98%, the same experiment on HMDB51 performs poorly (35%). This indicates that it may be necessary to take more into account action dynamics, in accordance with psychophysical experiments in which both motion and shape are critical for visual recognition of human actions [110, 111]. On top of that, entirely forgetting the spatio-temporal structure produces paradoxes even in simple mental experiments: by scrambling around the frames in a spatio-temporal volume we can obtain videos with absolutely no meaning whatsoever to us, but with roughly the same descriptors as the original one. In fact, the temporal factor is somewhat implicitly taken into consideration, as descriptors are extracted at multiple space and time scales.

**Dynamical generative modelling.** Indeed, encoding the dynamics of videos or image sequences can be useful in situations in which it is critically discriminative. Moreover, actions of sometimes very different lengths have to be encoded in a homogeneous fashion in order to be compared (“time warping”). Encoding each video sequence by means of some sort of dynamical generative model has been proven effective when coping with time warping or temporal segmentation [51]. In these scenarios, action recognition reduces to classifying dynamical models. Hidden Markov models [23] have been indeed widely employed in action recognition [43, 51] and gait identification [54, 9]. HMM classification can happen, for instance, by evaluating the likelihood of a new sequence with respect to the learnt models, or by learning a new model for the test sequence through the Expectation-Maximisation (EM) algorithm [67], measuring its distance from the old models, and attributing to it the label of the closest one(s).

Researchers have proposed the use of linear [5], nonlinear [25], stochastic [42, 24] or even chaotic [1] dynamical systems. Chaudry et al [10], for instance, employ nonlinear dynamical systems (NLDS) to model times series of histograms of oriented optical flow, measuring distances between NLDS via Cauchy kernels. As we recalled above, graphical models allow us to address the issue of localization [56], can be used as a tool to infer “plots” of videos [28] and, in situations in which crowds are present (as in surveillance scenarios), are suitable to describe the crowd’s behaviour in a way similar to the physical fluids modelling.

---

**Proposal Part B**

---

**Figure 4.** The “baseline”, Bag-of-Feature algorithm for action classification [87].
Issues with generative modelling. As we also argued above, classical dynamical models can be too rigid to describe complex activities, or prone to over-fitting the existing training examples. In probabilistic graphical models, such as (for instance) hidden Markov models, a major cause of overfitting is that they need to estimate from the training data unique, “precise” probability distributions describing, say, the conditional probabilities in an MRF, or the transition probabilities between the states of a traditional or a hierarchical hidden Markov model. To remain in the HMM example, there are efficient ways of dealing with this estimation problem, involving, respectively, the Expectation-Maximization (EM, [67]) and the Viterbi algorithm [65]. When little training data are available, the resulting model will depend quite strongly on the prior assumptions (probabilities) about the behaviour of the dynamical model.

1.2.3 Contributions: Pushing the boundaries in generative and discriminative approaches

From our brief discussion it follows that: on one side, discriminative approaches, thought successful in limited experiments in controlled environment, need to be extended to include a description of the spatio-temporal structure of an action, if they are to tackle issues such as action segmentation/localization, multi-agent activities, and the classification of more complex activities. On the other hand, generative graphical models have attractive features in terms of automatic segmentation, localization and extraction of plots from videos, but suffer from a tendency to overfit the available, limited training data. On top of that, more advanced techniques for the classification of generative models are necessary to cope with inherent variability and presence of covariates.

With this project we propose to break new ground in all these respects, with significant impact on the real world deployment of action recognition tools, by designing novel modelling techniques (both generative and discriminative) able to incorporate the spatio-temporal structure of the data, while allowing for the necessary flexibility induced by a generally limited amount of training information.

Introducing structure in discriminative models. In the field of discriminative modelling, we plan to build on recent progresses on the use of part-based discriminative models for the detection of 2D objects in cluttered images. If we think of actions (and even more so for complex activities) as spatio-temporal “objects”, composed by distinct but coordinated “parts” (elementary motions, simple actions), the notion of generalizing part based models originally designed for 2D object detection to actions becomes natural and appealing. In particular, as it is the case for objects, discriminative action parts can be learned in the framework of Multiple Instance Learning (MIL) [126, 127]. Consider a one-versus-all classification problem. In MIL, a discriminative model is learned starting from a bag of negative (of the wrong class) examples, and a bag of examples some of which are positive (of the right class) and some are negative (but we do not know which ones). Think, in our case, of all possible spatio-temporal sub-volumes in a video sequence within which, we know, a positive example of a certain action category is indeed present (but we do not know where). An initial “positive” model is learned by assuming that all examples in the positive bag are indeed positive (all sub-volumes of the sequence do contain the action at hand), while a negative one is learned from the examples in the negative bag (videos labelled with a different action category). Initial models are updated in an iterative process: eventually, only the most discriminative examples in the initial, positive bag are retained as positive. A flexible constellation of the most discriminative “action parts” can then be built from the training data to take into account the spatio-temporal structure of the activity at hand. Such an approach builds on the already significant results of discriminative models, but addresses at the same time several of the challenges we isolated in our analysis of the state of the art: 1 – complex activities can be learned and discriminated; 2 – localization in both space and time becomes an integral part of the recognition process; 3 – multi-agent action recognition now becomes standard practise as the presence of more than one action is assumed by default.

Move towards imprecise-probabilistic generative models. As for generative modelling, addressing the issues of inherent variability (which causes data overfitting) and influence of the covariate factors (which make rigid classification techniques inadequate) requires, on one side, to move beyond classical, “precise” graphical models; on the other, to develop a theory of classification for generative models which allows for robustness and flexibility. We have seen above that classical graphical models require to estimate a number of probability distributions from the training data. If the training data form a small subset of the whole universe of examples (as it is always the case), the constraint of having to determine single probabilities necessarily leads to overfitting.

In opposition, imprecise-probabilistic models replace such single (“precise”) probability distributions by whole convex closed sets of them, or “credal sets” [81]. Graphical models which handle credal sets, or “credal networks” [66], are a promising way of solving the overfitting problem, as they allow the actual evidence provided by a necessarily limited training set to determine only a set of linear constraints on the
true value of the probability distributions to estimate. Despite a number of early successes [67], progress in this field has been hampered by the computational complexity of inference algorithms in such networks. This has made finding computationally efficient or even feasible counterparts of the classical EM and Viterbi algorithms rather difficult. Recently, however, significant progress has been made towards efficient exact inference algorithms in credal trees [62,63]. This research, initiated and actively pursued by members of the present consortium (SYSTeMS and IDSIA) has opened up the development of efficient algorithms for imprecise hidden Markov models [64], a significant special case of imprecise-probabilistic graphical model.

Imprecise probabilistic graphical models allow a more flexible definition of action model: an imprecise model is indeed equivalent to infinitely many classical ones. At the same time, they retain the desirable features of classical graphical models in terms of segmentation and concise description of the observed motion. In addition, recognition algorithms based on such models (when compared to “precise” classifiers) can return more than a single action label: empirical comparisons generally show that the imprecise algorithm returns more than a single output when the precise algorithm just returns a wrong output: an extremely desirable feature, especially for security applications.

**Novel classification techniques for (precise/imprecise) generative models.** In the generative approach, action classification reduces to the classification of (precise/imprecise) graphical models. In the “precise” case, a number of distance functions have been introduced in the past (e.g., [52]), and a vast literature about dissimilarity measures for Markov models also exists [22], mostly concerning variants of the Kullback-Leibler divergence [33]. However, as the same models (or sequences) can be endowed with different labels (e.g., action, ID, no single distance function can possibly outperform all the others in each and every classification problem. On top of that, the variation induced by the many nuisance factors) makes any naïve approach to the classification of generative models doomed to fail. A reasonable approach when possessing some training data is instead trying to learn a supervised fashion the “best” distance function between models for a specific classification problem, for instance by employing differential-geometric methods [2,13,34]. Manifolds of generative models have been studied, for instance in [16], while the idea of supervised learning of distance functions has been successfully proposed and applied in the past in various contexts [3,4,48]. The same holds for imprecise graphical models. Dissimilarity measures for imprecise models can be derived, for instance, from distances between convex sets of distributions (as proposed by IDSIA).

As graphical models are complex objects, classification frameworks based on the “structured learning” approach to SVM classification [69] have also the potential to deliver state of the art results. Manifold learning based approaches to the classification of generative models encoding complex activities subject to a number of nuisance factors can help tackle the crucial issue of the presence of a large number of covariate factors, pushing towards “recognition in the wild” scenarios [37].

1.3 S/T methodology and associated work plan

1.3.1 Breakthroughs

To summarize: in order to progress towards the actual real-world deployment of action recognition in the different potential scenarios, the following challenges need to be tackled: 1– localization in space and time (segmentation); 2 – presence of multiple actors; 3 – influence of covariance/nuisance factors which make unconstrained recognition “in the wild” so difficult; 4 – inherent variability of actions with the same meaning, and corresponding overfitting issues due to limited training sets; 5 – analysis of complex activities rather than elementary actions.

The underlying idea we support with this project is that these issues can be tackled only by properly modelling the spatio-temporal structure (which combine localization and dynamics) of the motions to analyze. We propose to do so in both the generative and the discriminative approach to recognition, in order to achieve the following breakthroughs, which reflect the main challenges described above:

**Breakthrough #1: action localization in space and time.** An often neglected step, localizing an action in both space (image region) and time is the first necessary step in any action recognition framework. Modelling the spatio-temporal structure of the action/activity is paramount to this purpose. Generative graphical models have been explicitly designed for these purposes; novel discriminative models which incorporate such a structure will be developed.

**Breakthrough #2: recognition in the presence of multiple actors.** If more than one person is present in the scene, their individual behavior might have to be analyzed. Once again, spatio-temporal discriminative models potentially allow us to analyze their behavioral pattern separately. The same holds for
multiple actions performed by the same person.

Breakthrough #3: presence of covariance factors and recognition "in the wild". In order to bring action recognition out of our research labs, and towards extensive deployment in the real world, the issue of the manifold nuisance factors that affect human motion has to be somehow addressed. Again, we propose two parallel research lines to tackle this problem, one based on imprecise-probabilistic graphical models, classified in a flexible, adaptive way by learning classifiers from the actual distribution of the data in the training set; the other, founded on flexible spatio-temporal discriminative models.

Breakthrough #4: inherent variability and overfitting due to the limited size of training sets. In a machine learning framework, the ability to classify (in our case, to recognize actions or activities) is acquired from a training set of (labelled or unlabelled) examples. Unfortunately, collecting (and labelling) a large training set is a long and costly procedure: fully processed training sets for which ground truth is provided (think action labels and the corresponding bounding boxes in the video sequence) are a scarce resource. On the other hand a massive amount of unlabelled data is theoretically available, but in practise cannot be used for supervised learning of motions. We propose to cope with the overfitting issue by introducing a more robust, flexible class of "imprecise-probabilistic" graphical models which allow for uncertainty in the determination of the probabilities which characterize the model.

Breakthrough #5: moving from elementary actions to complex activities. Finally, in order to learn and recognize more complex, sophisticated activities, we need to incorporate some sort of description of the activity itself: in this project we propose to do so by both developing new spatio-temporal discriminative models, and by moving to a more flexible class of "imprecise" stochastic generative models.

1.3.2 Overview of the S/T methodology and project articulation into Work Packages

To achieve such a goal, we articulate our project into three parallel lines of research, which amount to the development of novel, ground breaking methodologies in three distinct scientific fields: 1 – structured discriminative models, i.e., collections of discriminative models able to describe the spatio-temporal structure of a motion; 2 – a fully fledged theory of imprecise-probabilistic graphical models, capable of retaining the good features of traditional graphical models, while allowing more flexibility in the modelling the inherent variability of actions, especially considering the limited size of any realistic training set; 3 – a novel theory of classification for both precise and imprecise graphical models, able to cope with the large number of nuisance factors and push towards action recognition “in the wild”.

The project is split into six interdependent work packages (WPs), one for each of the three methodological developments, plus one dedicated to the necessary analysis of novel algorithms for the extraction of suitable features from video/range sequences, one explicitly devoted to the integration and validation in various scenarios of the algorithms developed in the three research pipelines, and one for project administration.

The goal of WP1 is the development of a theory of imprecise-probabilistic graphical models, in particular learning and inference techniques which generalize classical methods for models based on sharp, “precise” assessments. The first step is the identification of a reliable technique to learn imprecise models from incomplete data, via the generalisation of the classical Expectation-Maximization algorithm to the imprecise-probabilistic framework, and its specialisation to imprecise hidden Markov models (iHMMs). Later on these tools will be extended to general imprecise-probabilistic graphical models. This concerns both the case of directed models, for which the class of “credal networks” has already been formalised and studied, and the case of Markov Random Fields (MRFs), for which an extension to imprecise probabilities represents a truly groundbreaking research challenge. This will give more flexibility to the description of the motion’s dynamics (e.g., by relaxing the Markovian assumption to support more complex topologies).

The aim of WP2 is to develop a theory of classification for both imprecise-probabilistic and classical generative models, able to tackle the issue with nuisance factors. To classify a new instance of an action we need to compare the corresponding generative model with those already stored in a labelled training set. This requires the development of dissimilarity measures for generative models, both precise and imprecise. In particular we will investigate techniques for learning the most appropriate distance measure given a training set. In the second part of WP2, these measures will be applied to classification proper, by exploring the use of innovative frameworks such as structured learning and compressed sensing.

The aim of WP3 is to develop a novel class of discriminative models able to explicitly capture the spatio-temporal structure of human motions. We can decompose this task into two parts. First, we need to identify the “most discriminative” (in the sense that they contribute the most to discriminating the presence of an action of a certain class) parts of an action/activity: we propose to use Multiple Instance Learning (MIL) techniques to learn such action parts from a weakly labelled training set. The second task is to arrange these parts in a coherent spatio-temporal structure. We will build on recent advanced in the field of 2D object
detecting, paying attention to specific features of spatio-temporal action recognition.

In order to push the boundaries of action recognition, the need for acquiring more adequate test-beds arises. Gathering and creating a large corpus of labelled videos/range sequences to be treated as benchmarks for the tools we intend to develop is the purpose of the first half of WP4. This looks particularly cogent once we note how most current test-beds still focus on elementary actions, while the range data modality is still in its infancy. In addition, the study of the process of extracting salient feature measurements from monocular, stereo and range sequences is a crucial step in both generative and discriminative modeling.

The practical implementation of the newly developed methodologies as an all-purpose, comprehensive framework for action and activity recognition is the function of the final work package (WP5). The set of algorithms, together with the methodologies developed in WPs 1,2 and 3 will be validated in a number of important, real-world scenarios of which the consortium already has expertise: gait identification, gaming and entertainment, video retrieval, robotics, and facial expressions analysis. WP5 will have strong feedback links with each of the theoretical WPs and the feature selection/extraction process, which will take place throughout the entire lifespan of the project.

Finally, WP6 will concern the necessary management activities of the consortium, and the implementation of the impact/dissemination actions such as organization of special issues, workshops, press conferences. These actions are further detailed in Part 3 of the proposal.

1.3.3 Detailed description of the Work Packages

WP1 – Imprecise-probabilistic generative dynamical modelling

Task 1.1 – Imprecise EM (general theory)

In order to allow for a more robust generative modeling of actions we want to learn imprecise-probabilistic models from video sequences. Techniques have been indeed proposed, and are currently employed, to learn imprecise-probabilistic models from complete datasets. Yet, when coping with hidden variables (as is the case for HMMs) these approaches have to be extended to cope with missing data, a still open field of investigation. In particular, by extending what is done for traditional HMMs, we want to learn an imprecise probabilistic graphical model by means a generalisation of the EM algorithm. The latter is commonly used to learn (precise models) from incomplete datasets, due to the presence of latent variables. An imprecise probabilistic reformulation of both the maximum likelihood (frequentist) approach and the maximum a posteriori (Bayesian) approach to the EM algorithm can be considered and compared on synthetic data, in order to decide which one provides a more reliable estimate of the model probabilities. This task will be in turn splitted into the following subtasks:

SubTask 1.1.1 - Bayesian approach to imprecise EM

A sensible option in order to develop of an imprecise-probabilistic version of the EM algorithm consists on directly generalising the Bayesian formulation of the algorithm, in which the single uniform prior adopted there is replaced by a set of priors which express a condition of prior-ignorance about the process generating the data. By analogy with the case of complete data, we employ a collection of Dirichlet distributions and exploit the fact that the distribution is the conjugate prior of the multinomial likelihood.

SubTask 1.1.2 - Frequentist approach to the imprecise EM

As an alternative, a likelihood-based (i.e., frequentist) approach can be pursued in which the likelihood of the available (incomplete) data is evaluated according to any possible (precise) probabilistic model defined over the variables at hand. Then, the imprecise-probabilistic model summarizing all the precise models whose likelihood is over a certain (“alpha cut”) level is selected. While an analytical identification of these models is unfeasible, a recursive approach to the identification of an approximate estimate can be formulated.

SubTask 1.1.3 - Empirical comparison based on synthetic data

It is not obvious which one amongst the Bayesian-like and the frequentist-like approaches to imprecise EM is going to provide the most reliable estimates. We plan to generate synthetic data from randomly-generated probabilistic models, and compare the accuracy of the estimates obtained in the various cases.

Task 1.2 - Inference on Imprecise HMMs (iHMMs)

Hidden Markov models are a natural way of describing dynamical processes affected by uncertainty. Their popularity is due in part to the availability of efficient algorithms for doing inference with them: EM, calculation of probabilities and likelihoods and most probable explanation (using Viterbi-like algorithms) can be performed efficiently. The second goal of this work package is to promote the state of the art of inference algorithms for imprecise HMMs to the same level as those for the precise ones. This amounts to developing and implementing computationally efficient algorithms for the following basic inference tasks:

SubTask 1.2.1. - Lower and upper probabilities of state and state-observation sequences
Based on the general ideas and methods developed by the SYSTeMS-IDSIA teams and described recently in [62,63], we will derive formulae for the probabilities of such sequences in imprecise HMMs that lend themselves to backwards recursion, and therefore to generalisations of the backward algorithms to precise HMMs, essentially linear in the length of the Markov chain.

SubTask 1.2.2 - Computation of lower and upper likelihoods
These are the lower and upper probabilities of the observation sequences. Backwards recursion calculations for these entities should also be possible, based on the seminal work in [62,63].

SubTask 1.2.3 - Most probable explanation
The "most probable explanation" for precise HMMs consists on finding the state sequence (explanation) with the highest probability, conditional on the observation sequence. In an imprecise probabilities context, there are basically two ways of formulating this problem [94]. The "maximin" method is a pessimistic approach which seeks the state sequence with the highest lower probability, conditional on the observations. In the "maximality" method, one state sequence is considered to be a better explanation than another if it has a higher posterior probability in all the precise models that are consistent with a given imprecise model. Preliminary analyses [64] hint that the latter method is most likely to be able to be implemented efficiently.

Task 1.3 - Imprecise EM for hidden Markov models
Task 1.2 is about making inference on imprecise HMMs, however learnt. Yet, learning imprecise HMMs from video sequences is not trivial, as the number of operations to be performed for each iteration of the algorithm can grow exponentially with the input size. This task involves:

SubTask 1.3.1 - Specialisation of the EM algorithms to the HMM topology
The imprecise EM developed in Task 1.1 has to be first specialized to the special case of HMMs, based on the inference algorithms developed in Task 1.2.

SubTask 1.3.2 - Implementation of recursive formulae for parameter updating
Recursive formulae describing the revision of model parameters have to be determined.

SubTask 1.3.3 - Convergence and initialisation: empirical study
The convergence properties of the EM algorithm, specialised to the HMM topology, have to be assessed. Its robustness and the quality of the estimates with respect to parameter initialisation (a crucial issue even in the precise case) needs to be considered as well.

Task 1.4 - Imprecise Dynamical Graphical Models
HMMs are an example of probabilistic graphical models with a dynamic component, in a way the simplest example of dynamic Bayesian networks. A general theory for dynamic BNs is well established, but a similar comprehensive framework is still missing for credal networks, the imprecise-probabilistic generalisations of Bayesian networks. Contributing to the development of such a general theory is the final aim of this work package, paying special attention to higher-order HMM topologies able to provide a better model of the correlation between the frames of a video sequence depicting an action. We also intend to explore the possibility of modelling these relations by means of undirected graphs, providing a generalisation of random Markov fields (which are already widely used in computer vision) to the imprecise probabilistic framework.

SubTask 1.4.1 - Dynamical (and object-oriented) credal networks
A crucial step consist on investigating whether the ideas pioneered in [62,63] for inference in credal trees can be extended to more general types of probabilistic graphical models. Methods will then have to be devised for (recursively) constructing joint models from the local imprecise-probabilistic models attached to the nodes of such credal networks.

SubTask 1.4.2 - Inference algorithms and complexity results for dynamic credal networks
Based on those (recursive) methods for constructing joints, recursive or message-passing algorithms need to be developed for updating these networks based on new observations, computing lower and upper probabilities, likelihoods, and most probable explanation.

SubTask 1.4.3 - Imprecise Markov random fields
Markov random fields are graphical models based on undirected graphs particularly suited for modelling pixel-to-pixel correlation in image analysis [85]. The extension of our inference techniques to imprecise Markov random fields is potentially of enormous impact and will be pursued towards the end of WP1.

WP2 – Classification of generative dynamical models
Once represented video sequences as either precise or imprecise-probabilistic graphical models (for instance of the class of imprecise hidden Markov models), gesture and action recognition reduce to classifying such generative models. Different competing approaches to this issue can be foreseen.

Task 2.1 – Dissimilarity measures for imprecise graphical models
SubTask 2.1.1 - Modelling similarity between sets of probability distributions
A first sensible option is to consider imprecise graphical models, and iHMMs in particular, as convex sets of probability distributions themselves. A distance/dissimilarity between two imprecise models is then just a special case of distance between two sets of distributions. We can then compute the latter as the vector of the distances between all the possible pairs of distributions in the two sets, and take the minimum or the maximum distance, or we can identify a single representative element in the credal set (e.g., the maximum entropy element or the center of mass), to which distance measures for single distributions can be applied. These ideas have already been formalised by OBU and IDSIA: yet the evaluation of such distances require to define an optimisation problem whose efficient solution still needs to be developed.

SubTask 2.1.2 - Specialisation to the class of imprecise hidden Markov models (iHMMs)
The results achieved in the previous task should later be specialised to the sets of (joint) probability distributions associated with an iHMM. The challenge, here, is to perform the distance evaluation efficiently, by exploiting the algorithms developed in Task 1.2.

SubTask 2.1.3 - Supervised learning of similarity measures for precise generative models
A second option is to learn in a supervised way the “best” metric for a given training set of models [3, 4, 48]. A natural optimisation criterion consists on maximising the classification performance achieved by the learnt metric, a problem which has elegant solutions in the case of linear mappings [50, 57]. An interesting tool is provided by the formalism of pullback metrics: if the models belong to a Riemannian manifold $M$, any diffeomorphism of itself (or “automorphism”) induces such a metric on $M$ (see Figure 5). By designing a suitable parameterized family of automorphisms we obtain a family of pullback metrics we can optimise on. OBU already has relevant experience in this sense [13]. The second goal of Task 2.1 will be the formulation of a general manifold learning framework for the classification of generative (precise) models, starting from hidden Markov models, later to be extended to more powerful classes of dynamical models (such as, for instance, semi-Markov [51], hierarchical or variable length [26] Markov models) able to describe complex activities rather than elementary actions.

Figure 5. Once encoded each training video sequence as a graphical model (for instance a hidden Markov model, left) we obtain a training set of such models in the appropriate model space $M$ (right). Any automorphism $F$ on $M$ induces a map of tangent vectors on $M$, which in turn induces a pullback metric.

SubTask 2.1.4 - Supervised learning of similarity measures for imprecise-probabilistic models
Most relevant to the current project is the extension of metric learning approaches to imprecise HMMs first, and other classes of imprecise graphical models later. We will be able to use the results of SubTasks 2.1.1 and 2.1.2 to define a base metric/similarity for iHMMs, on which to apply our metric optimisation approach.

Task 2.2 - Classification of imprecise graphical models
Task 2.2 is designed to use a-apriori or learnt distance measures (developed in Task 2.1) to classify imprecise models (2.2.1) and imprecise HMMs in particular (2.2.2), but also to pursue alternative classification strategies based on kernel SVMs in an approach called “structured learning” (2.2.3), or adopt large scale classification techniques based on randomised projections onto a reduced, measurement space (2.2.4).

SubTask 2.2.1 - Distance-based clustering and classification of imprecise models
Distance functions between convex sets of probabilities (2.1.1) can be used in distance-based classifiers to classify arbitrary imprecise models representing credal sets. Such an approach can be extended to more general classes of imprecise graphical models.

SubTask 2.2.2 - Efficient specialisation to the case of imprecise HMMs
The results of SubTask 2.1.2 can in the same way be used to deliver distance-based classifiers for imprecise Markov models as a relevant special case, given the importance of HMMs in action recognition. A computationally efficient specialisation of these techniques to the HMM topology is sought.

SubTask 2.2.3 - Kernel SVM classification of imprecise models: a structured learning approach
Kernel SVM classification is a powerful tool, which allows us to find optimal implicit linear classifiers in sufficiently large feature spaces without having to compute feature maps explicitly, thanks to the so-called
“kernel trick”. “Structured learning” [89] is the sub-field of machine learning that deals with the set of techniques devoted to classifying complex objects, rather than simple collections of numbers (vectors), by learning how to map inputs to arbitrarily complex outputs which possess a possible complex internal structure [88]. This is clearly the case of graphical models, so that structured learning approaches can be proposed to classify stochastic models. The extension of SVMs to interval data able to cope with imprecise-probabilistic models will also be considered [144,145].

SubTask 2.2.4 - Compressed sensing applied to imprecise models

“Compressed sensing” [76] is concerned with projecting the available data onto a randomly picked subspace [71,72,73] (or measurement space) of the original space, provided the latter is of sufficiently large dimension [74,75]. OBU has recently started to explore the use of compressed sensing for large scale classification, common in image segmentation [85]. Indeed, the complexity of most available SVM solvers (in terms of both storage and classification time) prevents their application to truly large scale, real-world problems. Results proving that SVM classification can be performed efficiently after random projection, greatly reducing its computational requirement, have recently started to appeared. OBU has recently proposed (by making use of specific random projectors developed in the PSH context to kernelize hash functions [77]) a framework in which linear SVM classification equivalent to kernel SVM classification is performed in an implicitly defined compressed feature space, extending to kernel methods recently proven results on “compressed” classifiers, and dramatically reducing the computational complexity. We intend to bring to full development these techniques and apply them to the classification of imprecise graphical models.

WP3 – Dynamical discriminative modeling

As we argued above, not only localizing an action within a larger video is a difficult problem (as the spatial and temporal extent of an action is ambiguous even for a human observer), but discarding the structure of human motions can be detrimental to recognition, especially when dealing with complex actions. Actions can better be described as constellations of elementary parts: locating them and building separate discriminative models for each single part can much improve recognition rates, when compared to current BoF approaches which create statistical models of whole videos which may contain similar sub-motions, often causing confusion between action classes. We propose to learn discriminative action parts in a weakly supervised setting, taking action localization to massive datasets in which manual annotation would be very difficult [118], and develop action models suitable for classification and retrieval.

Task 3.1 – Multiple instance learning of discriminative action parts

The first step in the development of such a framework consists on learning discriminative models for the elementary action parts that compose the activity to describe/recognize. Manually annotating the location of space-time actions in recent datasets (e.g. [104] which contains over 6000 video clips) is a highly laborious and tedious task. Lacking annotated ground truth, we can still learn action parts in a weakly supervised setting via Multiple Instance Learning (MIL).

Let an action-part be defined as a BoF model bounded in a space-time cube (see Figure 7). Unlike previous BoF approaches which either generate one histogram per video clip [106], or combine histograms extracted from a spatial grid [87], we propose to represent each action-part as a histogram of visual words. In action classification datasets, an action class is assigned to each video clip, assuming that one action occurs in each clip. Learning action parts may then be viewed as a weakly labelled scenario, in which it is known that a positive example of the action exists within the clip, but its exact location is unknown.

Figure 6. In Multiple Instance Learning, positive bags (sequences) are those which contain at least a positive example (left, in blue). In our case, the examples of a bag (video sequence) are all its spatio-temporal sub-volumes (middle, in various colors). Right: plot of Dense Trajectory Features [117] for a real sequence from Weizmann, with positive (green) and negative (purple) examples generated by MIL.

The task can be cast in a Multiple Instance/Latent Support Vector Machine learning framework, in which the
training set consists of a set of “bags” (the training sequences), containing a number of BoF models (or “examples”, in our case SVM classifiers learnt for each sub-volume of the spatio-temporal sequence), and the corresponding ground truth class labels. The class label for each bag (sequence) is positive if there exists at least one positive example (sub-volume) in the bag (which ones are positive we initially do not know): see Figure 6. An initial “positive” model is learned by assuming that all examples in the positive bag are indeed positive (all sub-volumes of the sequence do contain the action at hand), while a negative model is learned from the examples in the negative bag (videos labelled with a different action category). Initial models are updated in an iterative process: eventually, only the most discriminative examples in each positive bag are retained as positive.

A MIL/Latent-SVM framework has been used in [116], where possible object part bounding box locations were cast as latent variables. Such an approach permits the self-adjustment of the positive ground truth data, better aligning the learned objects filters during training. In action detection, Hu et al.’s SMILE-SVM focuses on the detection of 2D action boxes, and requires the approximate labelling of the frames (and human heads) in which the actions occur. In contrast, we propose to cast 3D bounding structures as latent variables. When using standard SVMs in conjunction with BoF as single part models, MIL becomes a semi-convex optimisation problem, for which Andrews et al. have proposed two heuristic approaches [126], of which the mixed-integer formulation suits the task of action localisation.

**Task 3.2 – Learning and classifying structured discriminative models of actions**

Once the most discriminative action parts are learnt via the application of MIL to the “bags” corresponding to the given training spatio-temporal sequences, we can construct tree-like ensembles of action parts (as in the pictorial structure model [115] illustrated in Fig. 7) to be later used for both localizing and classifying actions/activities.

Felzenszwalb and Huttenlocher [125] have shown (in the object detection problem) that if the pictorial structure forms a star model, where each part is only connected to the root node, it is possible to compute the best match very efficiently. Generalizing the argument brought forward in [125], an action can be represented by an undirected graph in which the vertices represent the action parts. A cost function can be defined as a function of both the (BoF) appearance models of the individual parts and of the relative positions between pairs of action parts, whose maximization yields the best action configuration [125]. Felzenszwalb et al. have shown how the problem can be efficiently solved by dynamic programming.

**WP4 – Data collection and feature extraction**

**Task 4.1 – Novel benchmark datasets gathering**

A crucial step in reaching the overall goals of the project consists on validating the theoretical advances on the theory of imprecise models, classification, and discriminative modelling on challenging, real world data. The objective of WP4 is therefore to design and acquire a relevant collection of video sequences and set up a repository which allow for data organisation, upload and download by all partners. A number of datasets are already currently available. They can be classified in terms of the scenario they refer to, their significance in terms of number of action categories allowed and their size, their difficulty in terms of the nuisance factors considered, and the level of complexity of the motions captured (see Figure 8).

Action recognition datasets include the ViHasi, the LSCOM, the IRISA, the CMU RADAR and the Daily living activities datasets ([http://www.cs.rochester.edu/~rmessing/uradl/](http://www.cs.rochester.edu/~rmessing/uradl/)). Most interestingly, new testbeds have recently been acquired in order to tackle the problem of recognizing actions “in the wild”. The YouTube action database [37] contains 11 action categories: basketball shooting, biking/cycling, diving, etcetera ([http://server.cs.ucf.edu/~vision/projects/liujg/YouTube_Action_dataset.html](http://server.cs.ucf.edu/~vision/projects/liujg/YouTube_Action_dataset.html)). It is challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, etcetera. The HOHA (Hollywood) dataset has been collected in 2008 [87] and 2009 [39]: the second batch, in particular, is intended to provide a comprehensive benchmark for human action recognition in realistic and challenging settings ([http://www.irisa.fr/vista/Equipe/People/Laptev/download.html](http://www.irisa.fr/vista/Equipe/People/Laptev/download.html)). Some other aspects, though, such
as localization, retrieval and modality are still quite neglected.

Figure 8. Right: some, thought limited, datasets are already available in many of the selected scenarios for this project. From top-left to bottom-right: surveillance (Weizmann), sports (UCF Sports), recognition "in the wild" (YouTube and Hollywood action recognition datasets), in this order. Left: this interesting diagram shows how recognition rates have gone down when novel, more challenging datasets have been introduced: an effect clearly correlated with the number of action classes contemplated by these testbeds.

We will therefore focus on gathering new data in the scenarios for which the currently available datasets are judged insufficient or inadequate for our purposes. We will consider both synchronised video cameras and range (time of flight) cameras (in virtue of the impact they are having at this very moment). We are not aware of available action datasets based on range data.

SubTask 4.1.1 – Acquisition of test video beds from conventional cameras

As an example, most action databases assume a correct segmentation of the action to classify, preventing a proper testing of the localization performances of discriminative methodologies such as latent SVM, which have been designed to locate “action parts” in the first place. We will collect videos encompassing several different actions performed by one or more people, across different areas of the scene the acquired images. Another strong contribution of this project is a push towards the recognition of complex activities rather than simple actions. Most current datasets are just not adequate as testbeds in this sense. We will then gather a new state-of-the-art activity dataset, which will likely spur a host of subsequent research worldwide.

Finally, while a few datasets have pushed the boundary towards a larger number of action categories, and others have taken into consideration several nuisance factors, no datasets we are aware of feature a combination of both, a truly challenging experimental setting. We will work on filling this gap as well.

SubTask 4.1.2 - Acquisition of test beds from range cameras

Range cameras are changing the landscape of computer vision at this very moment, as they provide several advantages over stereo configurations for depth estimation. A time-of-flight (TOF) camera is a system that creates depth data via a principle similar to that of LIDAR. In contrast to stereo vision, where complex correlation algorithms have to be implemented, distance info is easy to extract from such an output. The most striking application of range cameras is perhaps Microsoft Kinect game console, which is already revolutionising the whole field of interactive video games (http://www.xbox.com/en-us/live/projectnatal/). OBU Vision Group enjoys continuing strong links with Microsoft Research, and has recently acquired range cameras to kick-start cutting-edge research in motion analysis. Range cameras are most suitable for indoor scenarios such as gaming (see WP5, Task 5.3). To our knowledge, no systematic acquisition of range sequence data for action recognition has been carried out yet: this will be a significant contribution of the project. In addition, DYN plans to film an actor through several Kinect PCs in order to acquire a number of sequences showing the actor performing real expressive scenes.

SubTask 4.1.3 - Data repository and dissemination among partners

As large amounts of data of different modalities will be acquired during the project in order to test the developed prototype in the different scenarios we selected, it is important to have a means to store and efficiently access this data for all participants, and for the public. This will involve the development of a secure web site, or an intranet section of the project website developed for wider dissemination purposes. To what extent the data can be made public will be discussed by the consortium as outlined in Section 3. Some legal issues could be involved, as the need for protection of personal data of the subjects involved will arise.
Task 4.2 – Discriminative feature selection and extraction

Prior to modelling actions, the video streams have to be processed to extract the most discriminative information: such “features” can be extracted frame by frame, or from the entire spatio-temporal volume which contains the action(s) of interest. Their selection and extraction are crucial stages of this project.

SubTask 4.2.1 - Feature extraction from conventional video

Even though, historically, silhouettes have been often used to encode the shape of the moving person for action recognition, their sensitivity to noise and the fact that they require (on conventional 2D cameras) solving the (inherently ill defined) background subtraction problem has caused their replacement with feature selection frameworks in which a large number of low level operators or filters are applied to images (or, for action, spatio-temporal volumes), while in a later stage the most discriminative ones are retained for recognition purposes. Interest point detectors such as the Harris3D [70], the Cuboid detector [119], and the Hessian detector [120] may be employed to extract salient video regions, around which a local descriptor may be calculated. A plethora of local video descriptors have been proposed for space-time volumes, mainly derived from their 2D counterparts: Cuboid [119], 3D-SIFT [121], HoGhoF [87], HOG3D [122], extended SURF [120], C2-shape features ([108], see Figure 9 left), and Local Trinary Patterns [123].

Figure 9. Sketches of the algorithms for C2-shape features (left) and Dense Trajectory Features (right).

More recently, Wang et al. [117] have proposed Dense Trajectory Features (Figure 9 right), a combination of HoG-HoF with optical flow vectors and motion boundary histograms. These dense features, combined with the standard BoF pipeline [107] (Figure 4), have been shown to outperform the most recent state-of-the-art Learned Hierarchical Invariant features by Le et al. [124] on challenging datasets. They are therefore strong initial candidate descriptors for space-time video blocks. The second objective of WP4 will be setting up an efficient feature selection/extraction framework implementing all current state-of-the-art feature extraction approaches, and build on them to developed in the course of this project novel techniques.

SubTask 4.2.2 - Feature extraction from range cameras

In the case of range cameras, the type of data requires a tailored feature extraction process [92,93]. For instance, in [91] methods for extracting stable features for camera and laser range finder for SLAM are discussed. These approaches are most relevant to the navigation scenario (see WP5, Task 5.5). In the indoor, gaming scenario C++ code is already available for robust silhouette extraction of multiple actors from depth maps provided for instance by the Kinect console. We will extensively test and improve these algorithms in the course of 4.2.2. In the virtual reality scenario, the RGB-Z signals capturing actor performances will be used by DYN and SUP to elaborate a 3D textured mesh along with its deformations related to specific expressions (facial gestures). The 3D model and face texture will be used to construct an Active Appearance Model: this RGB-Z AAM will then be used to track the faces during the sequences and associate each frame to a specific appearance vector, to be later used for recognition of facial gestures.

The fusion of range and traditional images will also be a goal of WP4.

WP5 – Integration and testing in different scenarios

Work Package 5 is the core of the project. As the parallel research “pipelines” concerning feature extraction from range and traditional cameras, dynamical generative modeling and structural discriminative modeling deliver their expected outcomes, and pieces of code are generated and tested within each individual subject area, the need for integrating all the code in a single coherent framework to be tested on the data gathered by OBU, SUP and DYN for the scenarios of choice during WP4 arises.

Task 5.1 – Integration in a single prototype

The three parallel pipelines in which the project is articulated contain individual test stages in which the generated specific algorithms are tested and validated to help the development process. However, a comprehensive validation stage in which the entire framework is first assembled and then tested on state-of-the-art action, activity and gait recognition databases is necessary to assess the smooth interaction between single components and the overall performance in terms of robustness and recognition rates.

Task 5.2 – Validation in the gait identification scenario

Originally driven by an initiative of US’s DARPA, identity recognition from gait has been proposed as a
novel behavioural biometric, based on recognizing people’s distinctive gait pattern. Despite its attractive features, though, gait identification is still far from being ready to be deployed in practice, as in real-world scenarios nuisance factors heavily affect recognition. Gait identification is basically video sequence classification, with a different label (identity). A number of gait datasets are currently available (others are not online anymore): the Southampton (http://www.gait.eecs.soton.ac.uk), the USF GaitID challenge (http://marathon.csee.usf.edu/GaitBaseline/) and the CASIA (http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp) databases are the major ones. Because of the sheer size and number of covariate factors the USF test-bed is considered state of the art. Our aim will be to obtain state-of-the-art results on all benchmark datasets, especially the USF GaitID one, which is already in possession of OBU.

Task 5.3 – Validation in the gaming scenario
In the console game scenario, action recognition can be seen as a way to “spice up” the gaming experience, currently based, for the most successful products, on the estimation of the pose of the player, and its subsequent use in games which reproduce physical activities such as sports (crossbow, tennis, bowling) and dance. “Fun” games can be imagined in which the player has to interact via gestures with synthetic characters: think for instance of “adventure” games. In this scenario, we plan to test our prototype on the data gathered by the consortium, and OBU in particular, with a focus on range/depth data for its potentially dramatic contribution to the success of the project and its wider impact in both academic and industrial communities. OBU’s links with Sony and VICON will be extremely valuable in this respect.

Task 5.4 – Validation in the video retrieval scenario
The commercial potential of an effective algorithm for video-mining cannot be possibly overstated. The problem has some original features with respect to other application of video analysis: first some sort of concise description of each video (perhaps in terms of short verbal description of their “plot”) has to be generated; then, this description has to be fed to data mining algorithms to retrieve videos with a “similar” description. The class granularity requested by video retrieval is huge: a programme has to be able to tell apart and retrieve footages whose description is different from others in sophisticated, subtle ways. Due to the limitation of the datasets currently available, and their focus on classification rather than retrieval, we plan to test our prototype on a proprietary testbed acquired in the course of the project.

Task 5.5 – Validation in the autonomous navigation scenario
An autonomous navigation scenario, in which a robotic vehicle must be able to understand the behaviour of pedestrians on the road, and receive gestural commands in case of emergency, is at the centre of the new Intelligent Transport Systems Doctoral Training Programme established at OBU School of Technology. The research involved requires solving a number of sophisticated vision problems, such as the recognition of static (signs, cars) and moving (people, animals, bicycles) obstacles along the vehicle’s path or in its vicinity, which are quite related to the current project. Infrastructure and equipment are already in place in the form of an autonomous vehicle obtained from a quad bike (Figure 10 left), and a team working on it full time. The vehicle mounts stereo cameras, laser and will possibly mount a range camera as well. We will be able to exploit this infrastructure to test crowd behaviour analysis and gesture recognition in outdoor scenarios.

Task 5.6 – Validation in the virtual animation/facial gesture scenario
For this scenario we will focus on range camera data. The entertainment and advertising industry is interested in cloning actors, e.g. for Digital SFX movies such as “Avatar” or “The Curious case of Benjamin Button”. In this scenario, we will use the 3D model of the face developed in SubTask 4.2.2 to create a realistic Digital Double of the actor. The goal is to use such a clone to perform a realistic animation: thanks to the work developed in WP1, WP2 and WP3 it is possible to reproduce the style of the actor even if the animation is specified by key expression frames. As extremely realistic animations are needed here, DYN will add blendshape modelisation and the use of specific textures for the wrinkles to their rigging tools. The animations will be customised for a specific actor: in the expressive space of the actor (as developed by SUP), it is feasible to define specific trajectories in order to render accurate person specific animations. SUP will integrate the range information coming from several range cameras into its award winning [131] 2.5D Active Appearance Model (AAM, see Figure 10) [131] to detect and track facial features in real time, building on the multi-objective genetic algorithm developed in [129]. Starting from the work Supélec recently conducted on expression modelization, we plan to extract manifolds of facial expressions, called “Expressive Spaces” (Figure 10), tuned to the person analyzed. In the course of WP5 SUP will build on these results (2.5D AAM, multi-objective optimization, 3D face modelization, expressive spaces), in connection with the work developed in WP2 and WP4, to develop a fully fledged facial gesture framework. The challenge is to learn and be able to classify dynamic trajectories of an expressive face. This is highly related to Task 3.2 dedicated to the study of structured discriminative models of actions: an action in this case is the evolution of an expression. We want to go beyond static images, towards modelling the
expression’s dynamics to quantitatively represent the “style” of a specific actor.

Figure 10: The autonomous quod-bike (left) in possession of OBU. (Middle) In our “expression space” each expression is a point whose coordinates are the parameters of a 2.5D AAM. (Right) From the appearance space, we can extract a manifold which is the expressive space of the actor (the colored sphere), which can be used to synthesise the expression of a 3D character: each color is associated with an expression.

WP6 – Management and dissemination

The overall coordination of the project will be carried out by Dr Fabio Cuzzolin of OBU. He will be assisted by an administrative manager which will be hired at 25% of his/her time to assist with project management services and dissemination activities. The second objective of this work package is to ensure dissemination of the scientific results, as well as dissemination and exploitation in the industrial domain.

Task 6.1 – Coordination

Coordination comprises the organisation of a kick-off meeting with clear instructions on the administrative and financial responsibilities of the partners, reporting formats and time-tables. In subsequent Steering Committee meetings this will be a recurrent item on the agenda, next to the monitoring of technical progress and results (deliverables, milestones). Coordination further includes the preparation of annual progress reports and final reporting to the EC, including financial statements and audit certificates. The reports will be presented as drafts at the Steering Committee meetings where they will be approved before submission to the Commission. The reporting will be done according to FP7 rules. See Section 2 for more details.

Task 6.2 – Dissemination

The project foresees a wide spectrum of dissemination activities (see Section 3). In particular:

- targeted seminars to be held in companies active in the fields of gesture and identity recognition, human-machine interfacing, gaming, and biometrics; the existing links between the consortium and multinational companies such as Sony, Microsoft Research Europe and Google will be exploited;
- organisation of special sessions at major vision, machine learning and imprecise probability conferences, and special issues of international journals;
- media press conferences to present the results of the project to the wider public. An inspiring example is the project of the Laboratorio di Intelligenza Artificiale e Robotica of Politecnico di Milano on autonomous wheelchairs (http://airlab.elet.polimi.it/images/c/cb/LurchThesisCeriani.pdf), which has enjoyed several passages on national television networks, with a significant positive impact on the widespread diffusion of the project's outcomes;
- grand challenges on action and activity recognition, inspired by the successful VOC challenge;
- engaging government agencies potentially interested in security/surveillance applications of the developed robust recognition techniques;
- exploiting present and future academic links of the partners to kick-start, in case the results are as good as expected, follow-up research projects at national or European level, possibly large scale.

1.3.4 Milestones

The project's objectives stated in Section 1.1.3 are reflected in the following verifiable, measureable milestones (see the Table of page 27 and the PERT diagram in Section (iv)).

M1 – Development of efficient algorithms for inference and learning with imprecise HMMs. Achieved in month 18. Measured in terms of a software package in Matlab and C++ tested on synthetic and real data and relevant, high impact publications in AI/Imprecise Probability venues (see Section 3 for a list).

M2 – Full theory of learning and inference with general imprecise graphical models. Achieved in month 30. Measured in terms of a library of software packages on imprecise Bayesian Networks/MRFs tested on

synthetic and real data and relevant publications in AI/Imprecise Probability venues.

M3 – Dissimilarity measures b/w generative models both induced by credal sets and learned from the training set. Achieved in month 18. Measured in terms of both a library of routines in Matlab and C++ running and validated in real data, and publications in AI/Computer Vision venues.

M4 – Theory of classification of imprecise/precise generative models. Achieved in month 30. Measured in terms of a running library of Matlab/C++ routines implementing the different approaches to the problem (structured learning, k-NN, compressed SVM), made available to the public, and high impact publications in Machine Learning/ Vision journals and conferences.

M5 – Learning of discriminative action parts via MIL. Achieved in month 18. Measured by software libraries running and validated on real data, publications in first class vision/ML conference and journals.

M6 – Learning and classification of structured discriminative models. Achieved in month 30. Measured by software libraries running and validated on real data, available to the public, and publications in first class vision/ML conference and journals.

M7 – Dataset generation. Achieved in month 12. Extensive proprietary databases in different modalities such as monocular, synchronised and range cameras will be made public, via a dedicated web site.

M8 – State-of-the-art feature selection and extraction. Achieved in month 18. Measured in terms of a running library of Matlab/C++ routines implementing feature selections and extraction procedures in all modalities (monocular/spatio-temporal, stereo and range), tested in all scenarios with state-of-the-art results.

M9 – All purpose action recognition framework. Achieved in month 36. At project completion a toolbox collecting the routines implementing the different stages of the all-purpose action and activity recognition framework developed here will be made public in order to maximise the impact of this work.

M10 – Validation in all scenarios. Achieved in month 30. The framework and each individual component will be validated in all the relevant scenarios via thorough tests on all the public and proprietary databases. Success will be measured in terms of new state-of-the-art results on all the benchmark test-beds.

(i) Overall strategy of the work plan

Our workplan for the project is articulated into three, roughly parallel pipelines (see the PERT diagram of Section (v)). This is a natural consequence of the different tasks involved in delivering a robust action/activity recognition framework and prototype, in the variety of scenarios described above: 1- selecting and extracting robust features from video sequences in different modalities; 2- developing a coherent inference theory for imprecise-probabilistic graphical models; 3- working out adaptive ways of classifying dynamical generative models 4- introducing dynamics in state-of-the-art discriminative models.

The first area concerns the analysis of image sequences or the related spatio-temporal volumes, required to convert the video sequences into numerical information to be processed by the algorithms (WP4).

The second one encompasses the theoretical research focused on the development of the necessary learning and inference algorithms for the imprecise dynamic probabilistic graphical models we employ to describe actions and gestures (WP1), and the study the problem of classifying such complex objects in order to recognize actions, gestures or identities (WP2). Finally, WP3 is all about the development of novel dynamical discriminative models designed explicitly to represent and classify temporal patterns.

There is an obvious temporal relationship between workpackages 1, 2, 3 and 4, while the results of each pipeline are to be fed to the core of the project, the Integration and Validation work package WP5. The level of modularisation of the workload is designed to allow the five partners to work efficiently during the same period on the project, with limited risk of overlap or downtime.

All three pipelines are supposed to start from month 1, and run in parallel, managed by different groups. In fact, the “classification” work package (WP2) partially depends on the results of the “imprecise inference” one (WP1), but it is also autonomous in the sense that it will start by studying the problem of classifying more “classical”, precise dynamical generative models to later move to the imprecise case.

Feedbacks between “Integration and Testing” (WP5) and each of the other work packages will take place throughout the duration of the whole project: initial approaches and angles on the separate tasks will have to be revised in the light of the results obtained by the demonstrator(s).

These feedback loops will run for the entire second half of the project.

The number of research work packages (5) is, we believe, appropriate for a project this scale. However, as it can be noticed from our previous description of the individual work packages, each WP is accurately detailed and broken down into tightly focussed tasks and sub-tasks. This thanks to the fact that the consortium already possesses extensive expertise in all the tasks involved, allowing us to plan in much detail the proposed effort, facilitating in this way the monitoring by the Commission. The milestones/control points for the project have been discussed in Section 1.3.4, and are summarised in the Table at page 27.
(ii) Timing of the work packages (WPs)

The project is intended to cover a period of three years. The Gantt chart below depicts the overall schedule, including the timing of the different work packages.

<table>
<thead>
<tr>
<th>WP1</th>
<th>WP2</th>
<th>WP3</th>
<th>WP4</th>
<th>WP5</th>
<th>WP6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2m per</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The tables below are meant to give a more detailed description of the participation of each partner to each work package, stressing the contribution of each individual member of staff to the WPs. Each entry describes to which work package each person is assigned, for each two-month period in the lifetime of the project.

**OBU**

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**IDSIA**

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SYSTeMS**

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SUP**

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DYN**

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Proposal Part B: page [23] of [67]
(iii) Detailed work description broken into work packages

**Work package list**

<table>
<thead>
<tr>
<th>Work package No</th>
<th>Work package title</th>
<th>Type of activity</th>
<th>Lead partic no.</th>
<th>Lead partic. short name</th>
<th>Person-months</th>
<th>Start month</th>
<th>End month</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP1</td>
<td>Imprecise-probabilistic generative dynamical modeling</td>
<td>RTD</td>
<td>3</td>
<td>SYSTeMS</td>
<td>80</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>WP2</td>
<td>Classification of generative dynamical models</td>
<td>RTD</td>
<td>2</td>
<td>IDSIA</td>
<td>72</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>WP3</td>
<td>Dynamical discriminative modeling</td>
<td>RTD</td>
<td>1</td>
<td>OBU</td>
<td>39</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>WP4</td>
<td>Data collection and feature extraction</td>
<td>RTD</td>
<td>4</td>
<td>SUP</td>
<td>42</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>WP5</td>
<td>Integration and testing in different scenarios</td>
<td>RTD</td>
<td>1</td>
<td>OBU</td>
<td>66</td>
<td>13</td>
<td>36</td>
</tr>
<tr>
<td>WP6</td>
<td>Management and dissemination</td>
<td>MGT</td>
<td>1</td>
<td>OBU</td>
<td>13</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>312</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

1 Workpackage number: WP 1 – WP n.
2 Please indicate one activity (main or only activity) per work package:
   RTD = Research and technological development; DEM = Demonstration; MGT = Management of the consortium
3 Number of the participant leading the work in this work package.
4 The total number of person-months allocated to each work package.
5 Measured in months from the project start date (month 1).

## List of Deliverables

<table>
<thead>
<tr>
<th>Del. no.</th>
<th>Deliverable name</th>
<th>WP no.</th>
<th>Nature</th>
<th>Dissemination level</th>
<th>Delivery date (proj. month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Report on imprecise EM</td>
<td>1</td>
<td>R</td>
<td>PP</td>
<td>6</td>
</tr>
<tr>
<td>6.1</td>
<td>Web site dedicated to the project</td>
<td>6</td>
<td>O</td>
<td>PU</td>
<td>6</td>
</tr>
<tr>
<td>2.1</td>
<td>Report on similarity metrics for imprecise models</td>
<td>2</td>
<td>R</td>
<td>PP</td>
<td>10</td>
</tr>
<tr>
<td>1.4</td>
<td>Report on specialization of imprecise EM to iHMMs</td>
<td>1</td>
<td>R</td>
<td>PP</td>
<td>12</td>
</tr>
<tr>
<td>2.2</td>
<td>Software implementation of similarity metric b/w imprecise models</td>
<td>2</td>
<td>P</td>
<td>PP</td>
<td>12</td>
</tr>
<tr>
<td>4.1</td>
<td>Proprietary action datasets of stereo sequences</td>
<td>4</td>
<td>O</td>
<td>PU</td>
<td>12</td>
</tr>
<tr>
<td>4.2</td>
<td>Proprietary action datasets of range sequences</td>
<td>4</td>
<td>O</td>
<td>PU</td>
<td>12</td>
</tr>
<tr>
<td>4.3</td>
<td>Data repository</td>
<td>4</td>
<td>O</td>
<td>PU</td>
<td>12</td>
</tr>
<tr>
<td>5.1</td>
<td>First prototype of the overall action and activity localization and recognition framework</td>
<td>5</td>
<td>P</td>
<td>PP</td>
<td>12</td>
</tr>
<tr>
<td>4.4</td>
<td>Library of routines for automatic robust discriminative feature extraction from monocular videos and synchronised 2D frames</td>
<td>4</td>
<td>P</td>
<td>PP</td>
<td>15</td>
</tr>
<tr>
<td>2.4</td>
<td>Report on metric learning techniques for both precise and imprecise dynamical models</td>
<td>2</td>
<td>R</td>
<td>PP</td>
<td>16</td>
</tr>
<tr>
<td>3.1</td>
<td>Report on MIL framework for action part learning</td>
<td>3</td>
<td>R</td>
<td>PP</td>
<td>16</td>
</tr>
<tr>
<td>1.2</td>
<td>Report in inference algorithms for iHMMs</td>
<td>1</td>
<td>R</td>
<td>PP</td>
<td>18</td>
</tr>
<tr>
<td>1.3</td>
<td>Software prototype of EM and inference algorithms</td>
<td>1</td>
<td>P</td>
<td>PP</td>
<td>18</td>
</tr>
<tr>
<td>1.5</td>
<td>Software prototype of imprecise EM for iHMMs</td>
<td>1</td>
<td>P</td>
<td>PP</td>
<td>18</td>
</tr>
<tr>
<td>2.3</td>
<td>Report on classification of imprecise graphical models via similarity measures</td>
<td>2</td>
<td>R</td>
<td>PP</td>
<td>18</td>
</tr>
<tr>
<td>2.5</td>
<td>Software package on metric learning techniques</td>
<td>2</td>
<td>P</td>
<td>PP</td>
<td>18</td>
</tr>
<tr>
<td>3.2</td>
<td>Software prototype of MIL framework</td>
<td>3</td>
<td>P</td>
<td>PP</td>
<td>18</td>
</tr>
<tr>
<td>4.5</td>
<td>Library of routines for automatic robust discriminative feature extraction from range data</td>
<td>4</td>
<td>P</td>
<td>PP</td>
<td>18</td>
</tr>
<tr>
<td>6.2</td>
<td>Intermediate report to the EU</td>
<td>6</td>
<td>R</td>
<td>CO</td>
<td>18</td>
</tr>
<tr>
<td>3.3</td>
<td>Report on part based discriminative action models</td>
<td>3</td>
<td>R</td>
<td>PP</td>
<td>24</td>
</tr>
<tr>
<td>3.4</td>
<td>Software implementation of discriminative action models</td>
<td>3</td>
<td>P</td>
<td>PP</td>
<td>24</td>
</tr>
<tr>
<td>5.2</td>
<td>Preliminary report on classification/localization performances in all the available datasets and all scenarios</td>
<td>5</td>
<td>R</td>
<td>PP</td>
<td>24</td>
</tr>
<tr>
<td>6.3</td>
<td>Special session(s) on the project organized at major international conferences</td>
<td>6</td>
<td>O</td>
<td>PU</td>
<td>24</td>
</tr>
<tr>
<td>6.4</td>
<td>Special issue(s) on the project organized for major international journals</td>
<td>6</td>
<td>O</td>
<td>PU</td>
<td>24</td>
</tr>
<tr>
<td>2.6</td>
<td>Report on SVM classification of imprecise models</td>
<td>2</td>
<td>R</td>
<td>PP</td>
<td>28</td>
</tr>
</tbody>
</table>

---

6  Deliverable numbers in order of delivery dates. Please use the numbering convention <WP number>.<number of deliverable within that WP>. For example, deliverable 4.2 would be the second deliverable from work package 4.

7  Please indicate the nature of the deliverable using one of the following codes:

   R = Report, P = Prototype, D = Demonstrator, O = Other

8  Please indicate the dissemination level using one of the following codes:

   PU = Public
   PP = Restricted to other programme participants (including the Commission Services).
   RE = Restricted to a group specified by the consortium (including the Commission Services).
   CO = Confidential, only for members of the consortium (including the Commission Services).

9  Measured in months from the project start date (month 1).
<table>
<thead>
<tr>
<th></th>
<th>Activity Description</th>
<th>Page</th>
<th>Type</th>
<th>Prototype/Report</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>Report on general theory of imprecise-probabilistic graphical models</td>
<td>1</td>
<td>R</td>
<td>PP</td>
<td>30</td>
</tr>
<tr>
<td>2.7</td>
<td>Software package on SVM classification of imprecise models</td>
<td>2</td>
<td>P</td>
<td>PP</td>
<td>30</td>
</tr>
<tr>
<td>1.7</td>
<td>Software package on inference for imprecise-probabilistic graphical models</td>
<td>1</td>
<td>P</td>
<td>PP</td>
<td>32</td>
</tr>
<tr>
<td>1.8</td>
<td>Report on theory of imprecise Markov random fields</td>
<td>1</td>
<td>R</td>
<td>PP</td>
<td>34</td>
</tr>
<tr>
<td>2.8</td>
<td>Report on classification of imprecise models via compressed sensing techniques</td>
<td>2</td>
<td>R</td>
<td>PP</td>
<td>34</td>
</tr>
<tr>
<td>1.9</td>
<td>Software package with inference algorithms for iMRFs</td>
<td>1</td>
<td>P</td>
<td>PP</td>
<td>36</td>
</tr>
<tr>
<td>2.9</td>
<td>Software implementation of classification in compressed spaces</td>
<td>2</td>
<td>P</td>
<td>PP</td>
<td>36</td>
</tr>
<tr>
<td>5.3</td>
<td>Final prototype of the overall action and activity localization and recognition framework</td>
<td>5</td>
<td>D</td>
<td>CO</td>
<td>36</td>
</tr>
<tr>
<td>5.4</td>
<td>Final report on classification/localization performances in all the available datasets and all scenarios</td>
<td>5</td>
<td>R</td>
<td>PP</td>
<td>36</td>
</tr>
<tr>
<td>6.5</td>
<td>Final report to the EU</td>
<td>6</td>
<td>R</td>
<td>CO</td>
<td>36</td>
</tr>
</tbody>
</table>
List of Milestones

Milestones are control points where decisions are needed with regard to the next stage of the project. For example, a milestone may occur when a major result has been achieved, if its successful attainment is a required for the next phase of work. Another example would be a point when the consortium must decide which of several technologies to adopt for further development.

We detected the following milestones for our project:

<table>
<thead>
<tr>
<th>Milestone number</th>
<th>Milestone name</th>
<th>Workpackage(s) involved</th>
<th>Expected date</th>
<th>Means of verification</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Learning and inference with iHMMs</td>
<td>1</td>
<td>18</td>
<td>Software package on inference for imprecise HMMs released and validated on synthetic data</td>
</tr>
<tr>
<td>M2</td>
<td>Learning and inference with general imprecise graphical models</td>
<td>1</td>
<td>30</td>
<td>Software package on imprecise MRFs and BNs tested on real data</td>
</tr>
<tr>
<td>M3</td>
<td>Dissimilarity measures b/w generative models</td>
<td>2</td>
<td>18</td>
<td>Software implementation released and validated on datasets</td>
</tr>
<tr>
<td>M4</td>
<td>Classification of imprecise/precise generative models</td>
<td>2</td>
<td>30</td>
<td>Software implementation released and validated on datasets</td>
</tr>
<tr>
<td>M5</td>
<td>Learning of discriminative action parts via MIL</td>
<td>3</td>
<td>18</td>
<td>Software implementation released and validated on datasets</td>
</tr>
<tr>
<td>M6</td>
<td>Learning and classification of structured discriminative models</td>
<td>3</td>
<td>30</td>
<td>Software implementation released and validated on all datasets</td>
</tr>
<tr>
<td>M7</td>
<td>Dataset generation</td>
<td>4</td>
<td>12</td>
<td>Proprietary testbeds in multiple modalities acquired and released to the consortium</td>
</tr>
<tr>
<td>M8</td>
<td>Feature selection and extraction</td>
<td>4</td>
<td>18</td>
<td>Software implementation released and validated on all datasets</td>
</tr>
<tr>
<td>M9</td>
<td>All purpose action recognition framework</td>
<td>5</td>
<td>36</td>
<td>Integrated prototype assembled, validated and ready</td>
</tr>
<tr>
<td>M10</td>
<td>Validation in all scenarios</td>
<td>5</td>
<td>30</td>
<td>Prototype tested and validated on all available datasets, state-of-the-art results obtained</td>
</tr>
</tbody>
</table>

Their detailed description is given in Section 1.3.4. Their location within the project is shown in the PERT diagram of Section (iv), in terms of both duration in months and dependencies from the individual tasks the five research work packages are broken into.

---

10 Measured in months from the project start date (month 1).

11 Show how you will confirm that the milestone has been attained. Refer to indicators if appropriate. For example: a laboratory prototype completed and running flawlessly; software released and validated by a user group; field survey complete and data quality validated.
Work package description

<table>
<thead>
<tr>
<th>Work package number</th>
<th>Start date or starting event:</th>
<th>Work package title</th>
<th>Activity type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Month 1</td>
<td>Imprecise-probabilistic generative dynamical modelling</td>
<td>RTD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Participant number</th>
<th>Participant short name</th>
<th>Person-months per participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>IDSIA</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>SYSTeMS</td>
<td>44</td>
</tr>
</tbody>
</table>

Objectives

The goal of this WP is the development of a fully fledged theory of imprecise-probabilistic graphical models. This involves:
– the development of algorithms for learning imprecise-probabilistic models from incomplete datasets by generalizing the EM algorithm to the imprecise probabilistic framework;
– the development of efficient algorithms for imprecise hidden Markov models, by generalizing the classical forward and backward algorithms to the imprecise case;
– as a further step, the study of dynamic imprecise-probabilistic graphical models with more general topologies: in other words, the extension of Bayesian networks to imprecise probabilities.

Description of work

Task 1.1 – General theory of imprecise EM. The Bayesian approach (based on replacing single priors with sets of priors) and the frequentist approach (in which the likelihood of the available incomplete data according to any possible precise model is evaluated) to imprecise EM are developed and compared on synthetic data first.

Task 1.2 – Inference on imprecise HMMs. Backwards recursion calculations of lower and upper likelihoods and most probable explanation for imprecise HMM are studied, with the goal of promoting the state of the art of the algorithms for imprecise HMMs to the same level as those for the precise ones, in a computationally efficient way.

Task 1.3 - Imprecise EM for hidden Markov models. Learning imprecise HMMs from video sequences is not trivial. This task involves a specialisation of the EM algorithms to the HMM topology, the implementation of recursive formulae for parameter updating, and an empirical study of convergence and initialization properties.

Task 1.4 – Study of imprecise dynamical graphical models. This involves: investigating whether the ideas pioneered in [62,63] for inference in credal trees can be applied, or extended, to these more general types of probabilistic graphical models; developing recursive or message-passing algorithms for updating these networks and computing lower and upper probabilities and most probable explanations; exploring the possibility of similar extensions to imprecise Markov random fields.

Deliverables

D1.1 A report describing general theory of imprecise EM, algorithms and empirical analysis (month 6).
D1.2 A scientific report describing all the inference algorithms for iHMMs (month 18).
D1.3 A software prototype with the implementation of EM and inference algorithms (month 18).
D1.4 A report describing the specialisation of imprecise EM to hidden Markov models (month 12).
D1.5 A software prototype with its implementation (month 18).
D1.6 A scientific report describing the general theory of dynamic imprecise graphical models (month 30).
D1.7 A software package of inference algorithms for general imprecise graphical models (month 32).
D1.8 A scientific report describing the theory of imprecise Markov random fields (month 34).
D1.9 A software package implementing inference algorithms for iMRFs (month 36).

12 Please indicate one activity per work package: RTD = Research and technological development; DEM = Demonstration; MGT = Management of the consortium.
Work package description

<table>
<thead>
<tr>
<th>Work package number</th>
<th>2</th>
<th>Start date or starting event:</th>
<th>Month 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Classification of generative dynamical models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type 13</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Participant short name</td>
<td>OBU</td>
<td>IDSIA</td>
<td>SYSTeMS</td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>26</td>
<td>36</td>
<td>10</td>
</tr>
</tbody>
</table>

Objectives
The goal of this WP is the development of novel classification techniques for generative (dynamical) models, in both their precise (traditional) and imprecise versions. This involves:
– the study of similarity measures for both precise and imprecise-probabilistic dynamical models, by modeling similarity between sets of probability distributions (and specializing such measures to the case of imprecise HMMs), and developing techniques for the supervised learning of similarity measures for various classes of precise and imprecise dynamical models;
– the study of different alternative strategies for the classification of dynamical models, such as: distance based clustering using the developed similarity measures; kernel SVM classification of imprecise models in a structured learning framework; the application of compressed sensing techniques.

Description of work

Task 2.1 – Dissimilarity measures for imprecise graphical models. Different alternatives are explored. First option is to consider imprecise graphical models, and iHMMs in particular, as convex sets of distributions themselves. Results need to be specialised to the sets of (joint) probability associated with an iHMM. More sensibly, when possessing a training set of labelled video sequences/models we can learn in a supervised way the “best” similarity measure to adopt, by optimising classification performance over a parameterised search space of metric. The pullback metric framework of differential geometry can be exploited to this purpose. Such techniques can be generalized to imprecise-probabilistic graphical models, once we define a base metric/similarity for them.

Task 2.2 – Classification of imprecise graphical models. Distance functions between convex sets of probabilities (Task 2.1) can be used in a k-nearest-neighbor or other distance-based classifiers to classify arbitrary imprecise models representing credal sets. Different approaches to pursue in parallel are the use of kernels for imprecise-probabilistic models in a structured learning framework, and the exploration of compressed sensing techniques based on randomised projection to tackle the problem.

Deliverables

D2.1 A scientific report on similarity metrics for imprecise probabilistic models represented as convex sets of distributions, and their specialisation to the case of imprecise HMMs (month 10).
D2.2 Their efficient algorithmic implementation as a software prototype (month 12).
D2.3 A scientific report detailing the approach based on similarity measures to the classification of imprecise graphical models (month 18).
D2.4 A report describing supervised metric learning techniques for precise/imprecise dynamical generative models (month 16).
D2.5 Their efficient implementation as a software prototype (month 18).
D2.6 A scientific report detailing a supervised classification algorithm based on SVMs on imprecise models as complex objects in a structured learning approach (month 28).
D2.7 A software prototype with its implementation (month 30).
D2.8 A scientific report detailing the proposed compressed sensing approach to the classification of imprecise graphical models (month 34).
D2.9 A software prototype with its implementation (month 36).

13 Please indicate one activity per work package: RTD = Research and technological development; DEM = Demonstration; MGT = Management of the consortium.
Work package description

<table>
<thead>
<tr>
<th>Work package number</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start date or starting event:</td>
<td>Month 1</td>
</tr>
<tr>
<td>Work package title</td>
<td>Dynamical discriminative modeling</td>
</tr>
<tr>
<td>Activity type\textsuperscript{14}</td>
<td>RTD</td>
</tr>
<tr>
<td>Participant number</td>
<td>1</td>
</tr>
<tr>
<td>Participant short name</td>
<td>OBU</td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>39</td>
</tr>
</tbody>
</table>

**Objectives**

The goal of this WP is to develop a novel framework of discriminative models able to explicitly capture the spatio-temporal structure of human motions. This involves:

– identifying the most discriminative parts of an action or activity via Multiple Instance Learning (MIL) from a weakly labeled training set;

– learning and classifying pictorial structure models representing constellations of elementary, most discriminative parts.

**Description of work**

The work can be articulated into the following tasks:

**Task 3.1 – Multiple instance learning of discriminative action parts.** Let an action-part be defined as a Bag of Features model bounded in a space-time cube. The task here is to learn models for a set of most discriminative action parts, given a training set of video sequences in which an action class is assigned to each video clip, assuming that one action occurs in each clip. This is a weakly labeled scenario, where it is known that a positive example of the action exists within the clip, but the exact location of the action is unknown. The learning task can then be cast in a Multiple Instance Learning (MIL) framework, in which the training set consists of a set of “bags” (the training sequences), containing a number of BoF models (in our case SVM classifiers learnt for each sub-volume of the spatio-temporal sequence), and the corresponding ground truth class labels.

**Task 3.2 – Learning and classifying structured discriminative models of actions.** Once the most discriminative action parts are learnt by applying MIL to the “bags” corresponding to the given training spatio-temporal sequences, we can construct tree-like ensembles of action parts which can be later used for both localizing and efficiently classifying complex activities. A cost function can be defined as a function of both the appearance models of the individual parts and of the relative positions between pairs of action parts, whose maximization yields the best action configuration. The problem can be efficiently solved by dynamic programming.

**Deliverables**

**D3.1** A scientific report on the Multiple Instance Learning framework for the detection of the most discriminative action parts (month 16).

**D3.2** The efficient algorithmic implementation of the MIL framework as a software prototype (month 18).

**D3.3** A scientific report detailing the approach based on the optimization of constellations of elementary actions for the localization and classification of complex activities (month 24).

**D3.4** The efficient implementation of our novel framework for “structural” discriminative action modelling, localization and classification as a software prototype (month 24).

\textsuperscript{14} Please indicate one activity per work package:

RTD = Research and technological development; DEM = Demonstration; MGT = Management of the consortium.

Work package description

<table>
<thead>
<tr>
<th>Work package number</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start date or starting event:</td>
<td>Month 1</td>
</tr>
<tr>
<td>Work package title</td>
<td>Data collection and feature extraction</td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
</tr>
<tr>
<td>Participant number</td>
<td>1 4 5</td>
</tr>
<tr>
<td>Participant short name</td>
<td>OBU SUP DYN</td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>18 18 6</td>
</tr>
</tbody>
</table>

Objectives

The goal of this WP is to develop the infrastructure for the validation of the methodological breakthroughs developed in WP1-3 on the scenarios we isolated in WP5. To this purpose it is necessary to:

- define and acquire novel test-beds in multiple modalities: synchronised video sequences from multiple cameras (stereo), monocular videos, and range camera sequences; we will focus on complex activities and multiple actions taking place at different locations in the same video;
- create and manage a data repository for dissemination among the partners;
- design and implement the selection and extraction of salient pieces of information (features) in all the three modalities; features are later fed to parameter estimation algorithms (imprecise EM, WPs 1,2) or discriminative models (WP3) to locate and classify actions in video sequences. State-of-the-art feature selection algorithms are designed, in both 2D (e.g. dense trajectories) and 3D.

Description of work

**Task 4.1 – Novel benchmark datasets gathering.** Video sequence test-beds depicting multiple actors performing different actions and gestures at different spatio-temporal locations, with or without missing data and occlusions, under different environmental conditions, in different scenarios (human computer interaction, smart room, surveillance, identity recognition) are collected in three different modalities: stereo, monocular video, and range cameras. A data repository is created to store and efficiently share the gathered test-bed data among all participants, and eventually the public via a secure web site, or an intranet section of the project website developed for wider dissemination purposes.

**Task 4.2 – Discriminative feature selection and extraction** algorithms are designed starting from state-of-the-art approaches, such as dense trajectories in spatio-temporal volumes. Similar automatic feature selection algorithms are specifically designed for range data (depth maps). In this case reliable silhouettes or rough 3D volumetric representations of the moving person/body-part can be extracted. Unsupervised techniques for tracking volumetric features in an embedding space are also explored. Active Appearance Models of realistic animations are built via integration of range data by genetic algorithms.

Deliverables

**D4.1** A demonstrator including a corpus of video sequences acquired from multiple synchronised cameras, depicting multiple actors performing different actions/activities in different S/T locations in different scenarios (human computer interaction, smart room, surveillance, identity recognition) (month 12).

**D4.2** A demonstrator including a corpus of range data sequences, also picturing different people performing different gestures or facial expressions at the same time, mostly in indoor scenarios (virtual animations, gaming) due to inherent limitations of range cameras (month 12).

**D4.3** Shared data repository storing the gathered test-bed data in the three selected modalities, to be later made available to the wider academic community and the public (month 12).

**D4.4** A library of routines for automatic robust discriminative feature extraction from monocular videos and synchronised 2D frames (month 15).

**D4.5** A library of routines for automatic robust discriminative feature extraction from range data (sequences of depth maps) (month 18).

---

15 Please indicate one activity per work package: RTD = Research and technological development; DEM = Demonstration; MGT = Management of the consortium.
Work package description

<table>
<thead>
<tr>
<th>Work package number</th>
<th>Start date or starting event:</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Month 13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Work package title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integration and testing in different scenarios</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Activity type¹⁶</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Participant number</th>
<th>Participant short name</th>
<th>Person-months per participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OBU</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>SUP</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>DYN</td>
<td>30</td>
</tr>
</tbody>
</table>

Objectives

The goal of this WP is the integration of the code produced by the parallel pipelines on video analysis and feature extraction, inference and classification of imprecise-probabilistic graphical models and dynamical discriminative models into a single comprehensive prototype for action and activity recognition. In addition, we intend to validate the resulting prototype in all the relevant scenarios we detected: gait identification, gaming and entertainment, video retrieval, robotics, surveillance and virtual reality. We will test our approach on all the publicly available datasets on action, activity and gait recognition, in addition to the proprietary test-beds gathered by the consortium in all modalities.

Description of work

Task 5.1 – Integration in a single prototype. The algorithms for imprecise model parameter identification from partial data, inference on imprecise graphical models, discriminative modelling and feature extraction in multiple modalities are assembled in a coherent prototype. This is validated in multiple scenarios, and the results are fed back to WPs 1-4.

Task 5.2 – Validation in the gait identification scenario. The developed prototype is validated in the gait identification scenarios: the onus is on the presence of multiple covariate factors and monocular views.

Task 5.3 – Validation in the gaming scenario. In the gaming/console scenario, we plan to test our prototype on the data gathered by the consortium, and OBU in particular, with a focus on range/depth data and indoor scenarios, in which rough 3D silhouettes of the moving actors can be easily extracted.

Task 5.4 – Validation in the video retrieval scenario. In order to retrieve videos from a short description of their content, plots have to be automatically generated. A programme has to be able to tell apart and retrieve footages whose description is different from others in sophisticated, subtle ways.

Task 5.5 – Validation in the robotics/navigation scenario. In the robotics scenario, autonomous vehicles need to recognize what people around them are doing in order to prevent accidents, and are augmented with the capability of interacting with humans via gestures. Illumination/viewpoint factors are relevant.

Task 5.6 – Validation in the animation/facial expression scenario. We plan to integrate the information coming from range cameras in a 2.5D Active Appearance Model to track face features in real time, and to extract a manifold of facial expressions, called “Expressive Space” tuned to the person we analyze.

Deliverables

D5.1 A first prototype of the overall action and activity localization and recognition framework (month 12).
D5.2 A scientific report in which the results in terms of classification/localization performances in all the available datasets and all scenarios are thoroughly described (month 24).
D5.3 Final prototype of the overall action and activity localization and recognition framework (month 36).
D5.4 A scientific report in which the results in terms of classification/localization performances in all the available datasets and all scenarios are thoroughly described (month 36).

¹⁶ Please indicate one activity per work package:
RTD = Research and technological development; DEM = Demonstration; MGT = Management of the consortium.
# Work package description

<table>
<thead>
<tr>
<th>Work package number</th>
<th>Start date or starting event:</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Month 1</td>
</tr>
</tbody>
</table>

**Work package title**
Management and dissemination

**Activity type**
MGT

<table>
<thead>
<tr>
<th>Participant number</th>
<th>Participant short name</th>
<th>Person-months per participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OBU</td>
<td>13</td>
</tr>
</tbody>
</table>

## Objectives

The aims of this WP are the overall consortium management, both financially and in terms of coordination activities, and the dissemination of the results of the project at academic, industrial, and government level.

## Description of work

The work is articulated into the following tasks:

**Task 6.1 – Coordination.** This comprises the organisation of a kick off meeting with clear instructions on administrative and financial responsibilities of partners, reporting formats and time-tables, preparation of annual progress reports and final reporting to the EC, including financial statements and audit certificates. More details are given in Section 2.

**Task 6.2 – Dissemination.** This involves the spreading of the results and development of the project by means of targeted seminars, special sessions in major international conferences, special issues, grand challenges, media press conferences, personal engagement of intermediate bodies, agencies, and the government, exploitation of the consortium's networks of contacts to prepare the ground for follow up projects at national and European level. More details are given in Section 3.

## Deliverables

**D6.1** Web site dedicated to the project, with a public version open to the public/academic community, and an intranet version to allow an efficient exchange of ideas and resources between the partners (month 6).

**D6.2** Intermediate report to the EU (month 18).

**D6.3** Special session(s) on the project organized at major international conferences in imprecise probabilities, computer vision and artificial intelligence (month 24).

**D6.4** Special issue(s) on the project organized for major international journals in imprecise probabilities, computer vision and artificial intelligence (month 24).

**D6.5** Final report to the EU (month 36).

---

17 Please indicate one activity per work package:

RTD = Research and technological development; DEM = Demonstration; MGT = Management of the consortium.
Summary of effort

A summary of the effort is useful for the evaluators. Please indicate in the table number of person months over the whole duration of the planned work, for each work package by each participant. Identify the work-package leader for each WP by showing the relevant person-month figure **in bold**.

<table>
<thead>
<tr>
<th>Partic. no.</th>
<th>Partic. short name</th>
<th>WP1</th>
<th>WP2</th>
<th>WP3</th>
<th>WP4</th>
<th>WP5</th>
<th>WP6</th>
<th>Total person months</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OBU</td>
<td>24+2</td>
<td>36+3</td>
<td>18</td>
<td>30</td>
<td>9+4</td>
<td></td>
<td>126</td>
</tr>
<tr>
<td>2</td>
<td>IDSIA</td>
<td>36</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>72</td>
</tr>
<tr>
<td>3</td>
<td>SYSTeMS</td>
<td>44</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>54</td>
</tr>
<tr>
<td>4</td>
<td>SUPELEC</td>
<td></td>
<td></td>
<td>18</td>
<td>6</td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>DYN</td>
<td>6</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>36</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>80</td>
<td>72</td>
<td>39</td>
<td>42</td>
<td>66</td>
<td>13</td>
<td>312</td>
</tr>
</tbody>
</table>
iv) PERT diagram of the interdependencies between individual work packages

Figure 11. PERT diagram of the project. The causal dependencies (black arrows) between the five research work packages, broken down into their component tasks, are shown. The feedback links from the validations tasks T5.2-6 of WP5 and the methodological work packages WP1,2,3 are shown as red arrows. Milestones M1-M10 are denoted by circles: their location within the lifespan of the project is indicated in months.

The above PERT diagram illustrated the overall structure of the project in terms of causal dependencies between workpackages and their individual tasks, milestones, and temporal evolution. As we said the project is subdivided into four "pipelines": dynamic (classical and imprecise) generative modeling (WP1), generative model classification (WP2), dynamic discriminative modeling and classification (WP3) and feature extraction/selection (WP4), which reflect the main methodological contributions of the project. These four pipelines feed into the integration workpackage WP5 in which the developed techniques are tested in the different scenarios.

A number of points should be highlighted:

1. the individual work packages are subdivided into a number of very specific, detailed tasks;
2. each work package contains two milestones, roughly aligned with month 18 and month 36 (end of the project);
3. the project is highly parallelised, though dependencies exist between tasks in different work packages;
4. feedback links (represented by red arrows) are foreseen between the integration/testing stage of the project (in the different scenarios, T5.2-6) and the theoretical work packages WP1-3: as algorithms are assembled and tested, strategies and approach will likely have to undergo one or more cycles of revision.
v) Risk assessment and contingency plan

Some of the objectives of the proposals, reflected in the above work packages, admittedly involve a certain degree of risk.

In particular, the development of inference algorithms for imprecise hidden Markov models is cutting edge research at this time, while the formulation of a general theory for arbitrary imprecise-probabilistic graphical models poses a beautiful but serious challenge. The same can be said for the development of dynamical discriminative models, able to exploit the information provided by motion dynamics and couple it with the state-of-the-art performances of discriminative methods. Finally, the classification of “precise” graphical models is also subject of current, challenging research: this is in particular true for the theory of metric/manifold learning for dynamical models. Its extension to imprecise-probabilistic graphical models is an ambitious goal, which will be crucial to the success of the project.

These risks are balanced by the strong skills of the four academic partners in their respective area of expertise, and the specific features of the Coordinator with his interdisciplinarity expertise in both the applications object of this project (action, gesture and identity recognition) and its theoretical side (theory of imprecise probabilities and classification of dynamical models).

From a general point of view, as we mentioned above, the project is articulated into three different pipelines. Overall the project is designed as a “pincer movement”, in which different arms are designed to attack the problem from different angles in a complementary and cooperative manner. While a coherent formulation of inference and classification algorithms for imprecise-probabilistic graphical models is a challenge that cannot be met by simply building on existing results, WP4 activities (data gathering and feature extraction) are characterised by a lower level of risk, while, given OBU's expertise in computer vision applications and software development, we not envisage any insurmountable difficulty with integration and testing (WP5).

As for WP5's different scenarios, our strong industrial partner Dynamixyz has the necessary background to guarantee proper validation and testing in the virtual animation context, while OBU has running projects and extensive expertise in action recognition, gait identification, gaming and entertainment, feature selection and extraction, autonomous navigation and robotics.

Concerning the consortium's management, three of the four academic partners have smoothly and fruitfully worked together in recent years without any particular issue: it seems therefore reasonable to expect the same will happen in the course of this project as well.

To moderate risk, alternative approaches are considered from the start in the different workpackages, in order to ensure at least one of them will deliver within the desired time frame. For instance, in work package 2 we have foreseen to pursue several different alternative options to the classification of imprecise Markov and graphical models, based on clustering, structured kernel learning, and possible compressed sensing.

Further actions will be assessed by the steering committee in its periodic meetings and through informal contacts and video conferences between the partners. All actions deemed necessary will be proactively considered and enacted.
2.1 Management structure and procedures

The project involves partners from both industry and academic institutions with extensive experience in project management, though some of them contribute to a European project for the first time. In order to ensure clear responsibilities the following management structures will be installed.

2.1.1 Organisation

Overall structure

Figure 12. Overall management structure of the project.

Steering Committee The reasonable number of participants will likely make internal decision-making procedures quite simple. Nevertheless, a small Steering Committee will be necessary to ensure that the overall goal of the projects are met. The Committee will be in charge of monitoring the activities of the individual work packages and revising strategies in order to cope with unexpected developments, as part of our strategy to moderate the risk involved by some of critical tasks of the project.

It will be set up with one senior representative from each of the participants of the project, chaired by the Coordinator. All steering committee members will have the necessary authority within their organisation to
make decisions and commit staff and other resources in order to meet the needs of the project.

In particular, the steering committee will be responsible for:

- monitoring and reviewing the project progress and results with regard to deliverables and milestones;
- monitoring the activities of the work packages versus time and budget allocations;
- managing the integration of the software prototypes developed within the individual work packages and the consistency of the different project components;
- analysing Commission’s reviewers’ comments at a strategic level and initiating, in response, the necessary actions;
- making decisions concerning the update/revision of the scientific work and the implementation plan (mostly in correspondence of the milestones listed in Section 1.3.4), the scientific and administrative roadmaps, the allocation of funding and the acceptance of new parties (if deemed necessary);
- supporting the Coordinator in drawing up the reports and compiling the deliverables;
- drawing up very specific actions plans aimed at maximising the expected impacts, including in particular the organisation of special issues and workshops, contacting companies active in the fields involved, setting up press conferences, etcetera;
- the overall surveillance of dissemination activities.

The steering committee will meet at the beginning of the project and every six months thereafter, organised and chaired by the Coordinator. Extraordinary meetings can be conducted at the request of any consortium member.

The Coordinator will check whether such additional meetings can take place in the form of a telephone conference call.

Coordinator. The Coordinator (OBU) will be responsible for the overall management and will act as an intermediary between the Commission and the rest of the consortium (see Figure 12). He will be in charge of all legal, contractual, financial, and administrative aspects of consortium management and will supervise the project's administration.

In particular, the Coordinator will be responsible for:

- responding to important changes during the project lifetime and coordinating necessary adaptations to meet external requirements;
- transmitting the proposals of the work package leaders to the steering committee and managing the follow-up of their decisions;
- controlling manpower resources and expenditures;
- timely delivery of reports, deliverables, cost statements, financial audit certificates etc. to the European Commission;
- financial administration (including transfer of payments from the European Commission to the partners);
- distribution of any documents/information among the parties concerned, ensuring a constant, smooth communication flow;
- external affairs and presentation of the project (including public relations, press conferences, news articles, interviews); and
- overall follow-up, day to day management and conflict mediation between partners involved in the same work package.

The management office will be located in Oxford, UK. It will be led by the Coordinator Dr Fabio Cuzzolin. He will be assisted by a part-time administrative manager for the whole lifespan of the project.

Dr Fabio Cuzzolin has a considerable experience in grant applications, and he is already managing a small research group of two Ph.D. students, a PostDoc and a visiting student.

Dr Cuzzolin's recent internal proposal for a Ph.D. studentship in "Multi Sensor Fusion for Simultaneous Localization and Mapping on Autonomous Vehicles" has been funded under the "Intelligent Transport Systems" School doctoral training programme. A Ph.D. student has just started working on novel feature extraction techniques for action recognition, and the use of latent SVM classification. He has just been awarded a 122K GBP EPSRC First Grant, for a project entitled "Tensorial modeling of
FP7-ICT-2011-9
18/01/12 v1

STREP proposal

18/01/12 v1

dynamical systems for gait and activity recognition”. The project, which involves hiring a PostDoc for the second year, has received maximum marks by all four reviewers, and has started in August 2011.

His recent outline application for a Project Grant on ”The total probability theorem for finite random sets” with the Leverhulme Trust has passed the first stage, and the Trust has invited the submission of a full proposal on the topic. Invited full proposals have an acceptance rate around 50%-60%. This would involve an additional PostDoc for the three years of the project.

At European level, besides the present STREP, he is currently finalising as the Coordinator a Short Proposal for a Future and Emerging Technology FET-Open FP7 grant on “Large scale hybrid manifold learning”, with INRIA, Pompeu Fabra and Technion as partners. The project explicitly targets the use of compressed sensing for classification of large scale data (see WP2).

At national level he is setting up an EPSRC network on “Inference and decision making under uncertainty” together with J. Lawry (U. Bristol), J-B. Yang (Manchester Metropolitan), Frank Coolen (Durham), Jim Hall (Oxford University), W. Liu (U. Belfast) and others. He is involved as a partner in another EPSRC network proposal on Compressed Sensing, led by Mark Plumbley, Queen Mary University of London.

Most relevant to the current project, he is working on a Google Research Award on video retrieval under uncertainty with Professor T. Lukasiewicz (Oxford University).

Finally, he is in the process of submitting two separate EPSRC Project Grant proposals, one entitled “The total probability theorem for belief functions”, on the development of the imprecise-probabilistic approach for inference on time series; the other entitled “Locating and recognizing complex activities via dynamic generative/discriminative modelling”, in collaboration with Professor Philip H.S. Torr.

Oxford Brookes Vision Group. Dr Fabio Cuzzolin is a member of the wider OBU's Computer Vision Group (http://cms.brookes.ac.uk/research/visigroup/). The group comprises some twelve staff, students and PostDocs, and has a long and very successful history of grant management and collaborations with both academic and industrial partners. In recent years it managed a number of EPSRC and KTP projects worth around 2 million pounds. The group was mentioned four times in the UKCRC Submission to the EPSRC International Review Sep. 20064. It enjoys ongoing collaborations with companies such as 2d3, Vicon Life, Yotta, Microsoft Research Europe, Sharp Laboratories Europe, Sony Entertainments Europe.

The group’s work with the Oxford Metrics Group in a Knowledge Transfer Partnership 2005-9 won the National Best Knowledge Transfer Partnership of the year at the 2009 awards, sponsored by the Technology Strategy Board, selected out of several hundred projects.

Research and Business Development Office. The Coordinator will be assisted throughout the project by OBU's Research and Business Development Office (RBDO), with its long expertise in facilitating industrial collaborations and knowledge transfer.

Administrative manager. The proposed project is articulated enough to require a part-time administrative manager, who will assist the Coordinator in all administrative tasks. The administrative manager will also assist with the material organisation of the dissemination activities envisaged by the current project (while decision making will obviously reside with the Steering Committee).

Work package leaders (WPL). Each work package will be supervised by a Work Package Leader (WPL) to ensure the performance and progress of the work package with respect to the overall work plan. The WPL will summarise the progress reports delivered by the partners in the work package, and forward these to the Coordinator every three months for close monitoring. These reports will be discussed at the subsequent Steering Committee meetings. The progress reports will summarise the state of the work completed or planned. In the case of unforeseen events which require action, the reports will also describe what part of the work is behind schedule, and for what reasons. The work package leaders will be appointed by the partners according to their management skills and the scientific expertise required in the individual work packages. The responsibilities of the work package leaders are, in summary:

- monitoring the progress of the work package against time and budget allocations, ensuring that the work package fulfils the objectives defined by milestones and deliverables;
- alerting the Coordinator in case of delay or default in the performance of the work package;
- providing explanation for eventual delays/difficulties, and proposing appropriate actions in response;
- delivering quarterly management reports to the Coordinator;
- preparing proposals for an update of the work plan in case of unexpected developments; and

Proposal Part B: page [39] of [67]
preparing proposals for the inclusion of new parties, in case the opportunity for an extension of the consortium arise in the course of the project.

Work package members will mainly use electronic means, such as e-mail, phone and teleconferencing for their communication and coordination needs. If necessary, personal meetings will be organised by the work package leaders, to be combined as much as possible with other project activities.

Funding is requested for such physical coordination activities when deemed necessary.

Scale of the project. This is a relatively small, focused, yet ambitious project with very clear goals and objectives. The tools necessary to reach these goals have been clearly identified from the beginning, while risk moderation has been planned in the form of different research alternatives for the individual work packages.

This is reflected in a compact Steering Committee, whose limited scale will ensure greater flexibility. This will allow, for instance, the frequent use of conference calls as an efficient means of communication between partners. We believe this will increase the chances of the project of being steered flexibly and efficiently towards more promising or feasible strategies, in the perspective of ensuring that individual work programmes' and the whole project's targets are met.

As argued before, the number and choice of partners is a direct consequence of the objectives of the project and the tools we need to develop to reach them.

The well defined focus of the project in terms of goals and techniques is also reflected in the significant level of detail of the tasks individual work packages are subdivided into. The consortium knows what they want to accomplish, and they have a very clear idea of how to do that.

2.1.2 Communication strategy

The Coordinator will very regularly inform all the partners about the overall state of the project in terms of deliverables and milestones. This will be done by sending periodic newsletters (every three months).

In addition, a project website will be installed, articulated into a public and a non-public area, which will be accessible to project members only for communication purposes. In this way the partners will be able to share not only confidential data, but also ideas, problems, solutions.

A management handbook will be prepared, derived from experiences in other FP projects, in order to facilitate project management by defining a set of rules for the organisation of the day-to-day cooperative work. This will help to ease the management workload for all participants, and reduce overhead.
2.2 Individual participants

2.2.1 – OBU – Oxford Brookes University

Organisation. Oxford Brookes University, Oxford, UK is consistently recognised as the top modern university in the United Kingdom. The Department of Computing has had the second highest score among Brookes' departments in the last RAE assessment, ranking among the best 25 in the country. Academic staff include Prof. David Duce (co-chair of the Eurographics conferences), Prof. Rachel Harrison, Editor in Chief of Software Quality Journal, and Prof. Philip H.S. Torr, world leader in Computer Vision and Machine Learning. The School of Technology has established a doctoral training programme on “Intelligent Transport Systems” (http://tech.brookes.ac.uk/research/), whose infrastructure will be beneficial to this project.

Oxford Brookes Vision Group. Dr Fabio Cuzzolin is a member of the Oxford Brookes group founded by Professor Philip Torr (http://cms.brookes.ac.uk/research/visiongroup/), which comprises some twelve staff, students and post-docs who will add value to this project. Professor Torr was awarded the Marr Prize, the most prestigious prize in computer vision, in 1998. Members of the group, including the proposer, have recently received awards in 6 other conferences, including the best paper award at CVPR 2008, ECCV 2010 and BMVC 2010 and honorary mention at NIPS, the top machine learning conference. The group enjoys ongoing collaborations with companies such as 2d3, Vicon Life, Yotta, Microsoft Research Europe, Sharp Laboratories Europe, Sony Entertainments Europe. Its work with the Oxford Metrics Group won the National Best Knowledge Transfer Partnership of the year at the 2009 awards, selected out of several hundred projects. OBU has close links with Oxford University and Prof. Zisserman's Visual Geometry group, including a joint EPSRC grant and EU collaborations as well as Ph.D. students co-supervision.

Main tasks attributed and relevant experience. Oxford Brookes is the coordinator of the proposed project, and has a central role in its development. In particular, it contributes to the following work packages:
- WP2, a study of the supervised/unsupervised learning of similarity and distance measures between dynamical models, and the classification of stochastic dynamical models: this in virtue of Dr Cuzzolin's expertise on measuring distances between HMMs and other classes of models, and Oxford Brookes' vision group extensive expertise in SVM classification on Markov Random Fields;
- WP3 on the development of dynamical discriminative models based on part-based models whose parameters can be learned via Multiple Instance Learning;
- WP4 on video dataset gathering and image feature analysis, in virtue of the vision group's wealth of knowledge on automatic feature selection from images and spatio-temporal video volumes; and
- WP5 on the integration and testing of the different components of the framework, in virtue of the group's expertise in many of the highlighted application scenarios (gait, action recognition, gaming).

Profile of the people involved. The person in charge of the project for OBU is Dr Fabio Cuzzolin (Fabio.Cuzzolin@brookes.ac.uk). Dr Cuzzolin has been promoted to Reader in October 2011, and manages a small but growing research group. His research interests span gesture and action recognition, identity recognition from gait, and uncertainty modelling via non-additive or “imprecise” probabilities, to which he has contributed by developing an original geometric approach to belief functions and other uncertainty measures [12]. His scientific productivity is extremely high, as the thirty-six papers he has published in the last four years only attest. Dr Cuzzolin is currently author of 67 peer reviewed scientific publications, 59 of them as first or single author, including a monograph under review with Springer, two book chapters, 16 journal papers and 10 chapters in book series: http://cms.brookes.ac.uk/staff/FabioCuzzolin/. Eight other journal articles are in the process of being submitted in 2012. One of his papers won the best paper award at the recent Pacific Rim Conference on Artificial Intelligence Symposium 2008 [15]; another one was short-listed for best paper at the upcoming ECSQARU’11 conference. Recently, he received a best poster award at the recent ISIPTA’11 symposium on Imprecise Probabilities. Dr Cuzzolin is Associate Editor of IEEE Trans. SMC Part C, Guest Editor for Information Fusion, and collaborates with several other international journals in both computer vision and probability. He has served in the program committee of some 25 international conferences in both imprecise probabilities (e.g. ISIPTA, ECSQARU, BELIEF) and computer vision (e.g. VISAPP). He is reviewer for BMVC, ICCV and ECCV. He has supervised several MSc students, is currently supervising two Ph.D. students and is in the process of hiring a postdoctoral researcher.
The second partner involved in this project is the Imprecise Probability Group at the ‘Dalle Molle’ Insitute for Artificial Intelligence (IDSIA). IDSIA is situated near Lugano, a lakeside city in the Italian-speaking canton of Ticino in Switzerland.

**Organisation.** Founded in 1988, IDSIA is a non-profit oriented research institute for artificial intelligence, affiliated with both the Faculty of Informatics of the University of Lugano and the Department of Innovative Technologies of the University of Applied Sciences and Arts of Southern Switzerland. In Business Week's "X-Lab Survey" IDSIA was ranked in fourth place in the category "Computer Science - Biologically Inspired", above much larger institutions. IDSIA was also ranked in the world's top 10 of the broader category "Artificial Intelligence." IDSIA's academic staff includes more than 50 people. Director Prof. Luca Maria Gambardella has an international reputation for having invented and developed several ant colony optimisation algorithms to compute best-known solutions for sequential ordering, vehicle routing and travelling salesman problems. Director Jürgen Schmidhuber is well known for his work on machine learning, universal artificial intelligence, artificial neural networks, digital physics, and low-complexity art.

**IDSIA Group on Imprecise Probability.** Dr Alessandro Antonucci (alessandro@idsia.ch), the person in charge of the project for this partner, is a senior member of IDSIA's imprecise probability group founded and directed by Prof. Marco Zaffalon (www(idsia.ch/~zaffalon). The group includes a professor, four post-docs and two students, who can actively contribute to this project. Prof. Zaffalon, former President of SIPTA (Society for Imprecise Probability: Theory and Applications) and Area Editor for imprecise probabilities of the International Journal of Approximate Reasoning (IJAR), is one of the leading scientists in the field of imprecise probability. His group is one of the largest and most active units working on this field.

**Main tasks attributed and relevant experience.** IDSIA is a crucial partner for the proposed project. In particular, it contributes to the following work packages:

- WP1, involving the development of an imprecise version of the EM algorithm (to be specialised to the HMMs), in virtue of the long experience of the group in the learning of imprecise probabilistic graphical models from data (e.g., consider the seminal work of Prof. Zaffalon on the naïve credal classifier [101-103]) and possible developments of new approaches to this problem, recently considered by Dr Antonucci (i.e., the so-called "likelihood-based" approach to learning as described in [143] and [144]);

- WP2, which involves imprecise probabilistic graphical models and their classification, in virtue of Dr Antonucci’s strong background in the study and the application of graphical models and imprecise probability theories.

**Profile of the people involved.** Dr Antonucci’s research interests are mostly focused in the field of imprecise probabilistic graphical models. In particular, he is one of the main experts of credal networks (extensions of Bayesian networks to imprecise probability), from both a theoretical point of view and concerning applications. He is currently involved in a project granted by the Swiss Army for the application of credal networks to security problems, while an environmental network he developed was recently awarded by the local authorities of the canton of Ticino. He is the author of more than 25 peer reviewed scientific publications, including four journal papers and two chapters of which he is first author. He is the executive editor of the SIPTA Society (www.sipta.org) and was involved in the program committee of many international conferences on uncertain reasoning (UAI, ISIPTA, IPMU, ECSQARU, etc). He is reviewer for the JSTP and IJAR journals. He teaches numerical calculus at the faculty of informatics of the University of Applied Sciences and Arts of Southern Switzerland. He gave a number of public talks/lectures about imprecise-probabilistic graphical models, including a tutorial forum at AAAI-10 (one of the major conferences on AI) and the 'Summer School on Imprecise Probability' (in both 2010 and 2012). Additional information can be found on Dr Antonucci’s webpage: www.idsia.ch/~alessandro.
The third academic partner of the consortium supporting this proposal is the SYSTeMS research group at Ghent University, Belgium.

**Organisation.** Ghent University ranks among the largest universities in Belgium. Founded in 1817, it is a relatively young university. Today, after decades of uninterrupted growth, Ghent University is one of the leading institutions of higher education and research in the Low Countries. Located in Flanders, the Dutch-speaking part of Belgium and the cultural and economical heart of Europe, Ghent University is an active partner in national and international educational, scientific and industrial cooperation. It has more than 32000 bachelor and Master's students, and more than 2600 Ph.D. students. Ghent University employs around 7000 people. It scores consistently well in various international rankings, and is currently the only Belgian university in the top 100 of the Shanghai ranking.

**The SYSTeMS group.** Members of this group have research interests in various areas, such as stability and stabilisation, optimisation of dynamical systems, modelling and regulation of Lagrangian and Hamiltonian systems, machine learning, data mining, evolutionary optimisation, synchronisation of oscillators, large networks of interacting systems, control of discrete event systems, feedback control for on-line management of networks. An important unit in the group, involved in the present proposal, focuses on uncertainty modelling in dynamical systems, the foundations of uncertain reasoning, and imprecise probabilities for knowledge representation.

**Profile of the people involved.** The person in charge of the project for SYSTeMS will be Prof. Gert de Cooman, who heads the Probabilistic Inference unit.

He has made important contributions in a range of topics in the field:

- mathematical aspects of imprecise probability models;
- sets of desirable gambles as a foundation for uncertain reasoning;
- belief change with general information states, imprecise hierarchical uncertainty models, learning and optimisation using imprecise probability models, with applications to systems modelling, identification and control, classification;
- imprecise probability models for dealing with missing data [95];
- symmetry in imprecise probability models, and in particular issues of exchangeability;
- laws of large numbers and ergodic theorems for imprecise probability models;
- predictive inference and the Imprecise Dirichlet Model;
- applications of the Imprecise Dirichlet Model in game theory;
- epistemic irrelevance in credal networks, and efficient algorithms for inference in the resulting credal trees, and in particular hidden Markov models [62-64,97-100].

He has published more than 100 research papers on these topics in various scientific journals, books and conference proceedings, and is currently working on two books on the theoretical foundations of imprecise-probabilistic reasoning. He has been at the international forefront of research in his field, and is also a founding member and past president (2000-2006) of SIPTA, the international professional society devoted to research and education in, and dissemination of, the theory of imprecise probabilities. As such, he has played a very active role in organising most of the conferences and summer schools in the field. He has been a Visiting Fellow at Grey College (Durham University), and serves as a reviewer for, and member of the editorial board of, several journals.

**Main tasks attributed and relevant experience.** In this project, SYSTeMS will be mainly involved in the theoretical research focused on the development on inference algorithms for imprecise hidden Markov models. Considering his experience in, and knowledge about imprecise probabilities, and his pioneering work on imprecise credal trees and hidden Markov models, Gert de Cooman's group appears as a perfect partner to help achieve the goals set in the work packages in care of SYSTeMS.
The fourth partner of the consortium supporting this proposal is Supélec, short for "Ecole Supérieur d'Electricité", one of France's prestigious Grandes Ecoles.

Organisation. Supélec has a special position among France's leading engineering institutes, the reference in the fields of electric energy and information sciences. Internationally, with 440 engineers graduating each year, Supélec is on a par with the best departments of electrical and computer engineering of the top American or European universities. Supélec is an important training center for Ph.D students in the field of information, energy and systems sciences, and it welcomes more than 250 Ph.D students into its research programs and laboratory facilities. Supélec was founded in Paris in 1894. Today the school is spread over a three-campus network, located in Gif-sur-Yvette on the outskirts of Paris, Rennes in Brittany and Metz in Lorraine.

The school has 3 main missions: initial training of engineers, research and development, and professional training. Very often the presentation of research activities is stifled by a dilemma: how does one present at the same time a cross-disciplinary approach (ex. Systems Sciences) without betraying the specialists and their specific fields of activity? Nevertheless, Supélec's research is basically organized into seven major fields of study, all closely linked to Systems Sciences. Telecommunications is one of them: SCEE, based in Rennes, belongs to the latter.

SCEE (Signal, Communication & Embedded Electronics)'s objectives are the study and design of future communication systems based on Software Radio and Cognitive Radio concepts. The team has edited a book [137] which summarizes its three main research topics; among those, “Face Analysis and Synthesis” belongs to the “smart sensors” field. This line of research has been established more than fifteen years ago, and has since involved a small team of three to four persons: a permanent researcher with two or three Ph.D. students and a PostDoc. Face analysis for video communication coding was the main task during the first ten years: the objective was to model the face and its motion in order to create a parameterize model. Those parameters, rather than the 2D frame content compressed in a classical manner, were then transmitted to drastically reduce the bit rate. Face features detection was another important task.

Main tasks attributed and relevant experience. In this project Supélec will work on the scenario where an actor is in front of several RGB-Z cameras and plays a part in a scene. Supélec will be in charge of the 3D face modeling and tracking. As described in Section 2.3.1, the partner has proposed a multi-objective optimization framework for Active Appearance Models convergence from several cameras to track expressive faces. Supélec will build on the same framework and adapt such algorithms to the case of RGB-Z cameras, which we deem crucial to enhance the impact of the project (WP 4 and 5).

Their previous work on 3D face modeling will be used too in this project, in order to propose a real clone of the actor and deliver a realistic expressive clone.

Profile of the people involved. Renaud Séguier has been Professor at Supélec for fifteen years. After his PhD, he has received the HDR (Habilitation à Diriger des Recherches) degree in 2011 for his research activities in face analysis and synthesis. He has developed specific skills in spiking neurons analysis, with the purpose of detecting spatio-temporal signals, and has applied them to lip reading. He has developed a hybrid multi-objective genetic algorithm dedicated to face analysis in lip reading, and applied them to the optimization of several snakes to describe the mouth's contours. Later, he applied these techniques to pose estimation from several cameras. In the last eight years his work has mostly focused on the use of Active Appearance Models in face analysis. He has put forward several improvements to make these models more robust to illumination conditions, unknown expressions and identity. His students currently work on two French collaborative projects; in particular the group's research conducted during IMMEMO [133], in collaboration with ISIR (Institut des Systèmes Intelligents et de Robotique), has allowed him to win the Challenge dedicated to facial expression analysis at the last Face and Gesture conference in Action Unit detection [132].
2.2.5 - DYN – Dynamixyz

The industrial partner featured by our consortium is the French company Dynamixyz.

**Organisation.** Founded in April 2010, Dynamixyz is a simplified joint stock company (SAS – Société par Action Simplifiée) with headquarters located in Rennes (France). The company is specialized in high quality 3D face analysis and synthesis, for which there is growing demand and increasing potential for applications. With regard to analysis, for instance, few solutions are yet available capable of performing subtle detection of facial features in real-time. And yet the needs in this field are growing, as can be seen in the increasing use of reality applications at points of sale – virtual trying-on of glasses and virtual make-up are some examples – and in the marketing field’s requirements for detecting individuals’ gender and/or age. Synthesis is significantly growing, due to the widespread diffusion of video games and novel interactive entertainment experiences. There are still currently very few solutions for producing truly realistic facial animations – in most cases developing such animation can be costly and, depending on the application, the solutions offered are not always reliable. While more and more virtual humans are appearing on our screens, the current content creation solutions are not sufficient to meet the increasing level of realism demanded. Statistical systems which can produce good quality, on-the-fly animations are needed. Dynamixyz employs a large research team (six Ph.D holders specialized in the analysis/synthesis fields). The company aims to create high-tech products for professional users which can be easily integrated into any platform.

**Products.** Dynamixyz proposes two tools to animate and retarget the expression of an actor on a 3D character. “Director” [130] is a high quality facial animation tool designed for video game industry professionals and animation film production studios. Director can be used either as a plug-in for authoring – Director in 3DSMax and Motion Builder – or as an integrated library within a 2D/3D engine, for real-time use – Director RT. Supported by Dynamixyz technology, Director is capable of creating perfect facial animations from which highly realistic expressions can be generated, and to offer sophisticated lip syncing. “Performer” [131] is a marker-less facial performance capture solution which only requires a headset-mounted camera. The system can be used to animate 3D characters in films, video games or in any virtual environment for which realistic facial expressions are needed. Once the Performer head-set is in place on the actor’s head, a high accuracy computer vision system tracks the slightest of skin movements. Based on deformations in the skin’s texture, the system calculates the parameters required to animate the desired virtual character. The head-set mounted camera transfers video streams via a specifically designed wireless connection, leaving the actor completely free to play his/her role and to execute the movements they want the character to perform. Performer generates high quality output for film and video game production and an intermediate level of quality for applications involving real-time tasks (events, etc).

**Main tasks attributed and relevant experience.** The relevant experience of Dynamixyz is quite extensively summarized above in the product description of Director and Performer: the company can provide solutions to animate a 3D character and retarget the expression of an actor on a synthesized avatar. The work assigned to Dynamixyz for this project takes place in WP5: it consists on proposing a specific face rigging (skin and bones, blend-shapes, muscular models…) animatable from the Expressive Space. This manifold will be used by Dynamixyz for the retargeting. The realistic texture extracted in WP4 from the actor will be converted in a texture dedicated to synthesis. The partner will use specific rendering techniques coming from the video games industries (ambient occlusion, global illumination, subsurface scattering and soft shadows) to attain a realistic animated face.

**Profile of the people involved.** Olivier Aubault is a Ph.D engineer and the CTO of Dynamixyz. He is specialized in rendering techniques, and will allocate time for research activities and management. He participated in a national French collaborative project (RNRT V2NET) on 3D rendering of complex objects and in the Mpeg-4 group on mesh coding and animation. During his ten years at Orange Labs (French Telecommunication Company) he participated in many development projects in various domains, from real time rendering to 3D character animation.
2.3 Consortium as a whole

In addition to their individual scientific profiles, it is important to point out what the partners' previous expertise is specifically in relation to the project we propose and the tasks assigned to them. Indeed, the members of our consortium do already possess significant expertise in all the fields involved.

2.3.1 Past relevant results by the partners

Imprecise models. One of the main challenges of this project is to develop a complete theory (including learning, inference and classification algorithms) for imprecise-probabilistic graphical models. Work in this direction has been already done by SYSTeMS and IDSIA, which together with the (non-EU) group led by Fabio Cozman at the University of Sao Paolo in Brazil (http://sites.poli.usp.br/p/fabio.cozman/) can be regarded as the world leaders on research on credal networks. In particular, concerning dynamical probabilistic graphical models, both these groups have already considered an imprecise-probabilistic extension of hidden Markov models (iHMMs).

An inference algorithm for imprecise HMMS has been proposed in a joint work of SYSTeMS and IDSIA [62] which focuses on predictive inference: some ideas can nevertheless be extended to other, more general inference problems. This algorithm has been applied to character recognition, thanks to an already formalized learning procedure to obtain an imprecise HMM from complete data. Clearly, doing the same for the case of incomplete data (hidden variables) will prove more difficult, but can be possibly tackled via similar mathematical techniques. In another joint work by SYSTeMS and IDSIA [63], the same algorithm has been applied to multiple tracking problems [139]. Some ideas about learning these models from incomplete data have been already formalized by both SYSTeMS [142] and IDSIA [140], and also applied to the classification of time series [141].

As for the EM algorithm, papers by IDSIA have broken some ground on the problem of learning imprecise probabilistic graphical models from complete datasets; a number of SYSTeMS's and IDSIA's works deal instead with different approaches to the learning of general imprecise models from missing data [95,96].

In summary, a large body of papers by IDSIA and SYSTeMS [62-64,97-100] (as well as a recent joint work by IDSIA and OBU focused on action recognition) exist that can well be regarded as the strong foundation needed to develop the full theory of imprecise probabilistic graphical models envisaged by this project.

Generative modelling for human analysis, manifold learning. On its own, OBU has a significant track record of research in human motion analysis and recognition [18]. In particular, most relevant to the current proposal, the Coordinator has recently explored the use of bilinear and multi-linear models to identity recognition from gait [11,14], the promising branch of behavioural biometrics we considered as a scenario in WP5. In a strictly related topic, he is currently exploring manifold learning techniques for dynamical models representing (human) motions [13,16], in order to learn the optimal metric of the space they live in and maximize classification performance (compare WP2, Task T2.1). In the wider field of articulated motion analysis he has published several works on spectral motion capture techniques [17], focusing in particular on 3D surface point tracking and the consistent segmentation of body-parts in sequences of voxel-sets, as a preprocessing step to action recognition.

The wider Oxford Brookes vision group of which the Coordinator is a member is widely recognised as one of the top computer vision groups in the world, winner of numerous best paper awards and prizes, with an extensive expertise in object detection and segmentation, inference on conditional random fields (a class of probabilistic graphical models analogous to HMMs, but based on undirected graphs), and graph-cut algorithms for large scale image segmentation [85] (see http://cms.brookes.ac.uk/research/visiongroup for a complete list of publications).

The resources and skills provided by the group will be crucial to the success of this project.

In part-based discriminative modelling for video analysis, OBU is currently starting a multiannual project on action recognition in mobile robotics, as part of a doctoral training programme on Intelligent Transport Systems (http://mems.brookes.ac.uk/research/intelligent-transport-systems/). Different scenarios in which an autonomous vehicle has to recognize the behaviour of the street crowd, or individual people interact directly with the machine via gestural commands are explored, and contribute to the various scenarios envisaged in this project by WP5. One of the Coordinator’s Ph.D. students has just began to explore the application of part-based discriminative models inspired by the object detection problem to the localization and classification of actions, seen as collections of spatio-temporal parts.
**Scenarios of real-world deployment.** Strong expertise is also present in many of the manifold real-world application scenarios discussed in the introduction (and explicitly contemplated in WP5, see Section 1.3.3). The Coordinator has a running EPSRC (the UK Engineering and Physical Sciences Research Council) project on gait identification (http://cms.brookes.ac.uk/staff/FabioCuzzolin/projects/epsrc10first.html), while OBU's vision group has running collaborations with Sony and VICON on vision techniques for video games, in particular via range camera/Kinect technology. As just mentioned, doctoral work is being conducted on action recognition/crowd interpretation in the context of autonomous robotics. The Coordinator has a partnership with Oxford University's T. Lukasiewicz on a Google Research Award proposal on video retrieval/automatic plot generation under uncertainty.

Within **animation and facial expression analysis**, Supélec has worked on 3D face modeling [135] from classical RGB cameras in order to extract the 2D texture of a face and its shape at the same time. Stereo vision has been recently used to evaluate in real time the face pose and features on unknown persons. To that purpose, 2.5D Active Appearance Models (AAMs) have been proposed [128] and optimized in a multi-objective framework based on genetic algorithms. Supélec has recently won the Face and Gesture Challenge dedicated to expression detection [131], using its 2.5D Active Appearance Model adapted to unknown expressive faces. Dynamixyz, strongly linked to Supélec, has proposed a tool [130] to parameterize a 3D character. The proposed rigging was fair enough to allow a real-time animation of the proposed parameterized 3D character. It has also developed a tool [134] to retarget the expression of an actor wearing a head mounted camera to a 3D character. Dynamixyz is the only one company in the world able to provide such an instrument: the competitors (Image Metrics, Cubic Motion) sell the retargeting process as a service only.

### 2.3.2 How the participants constitute a consortium capable of achieving the objectives

The proposed project is a **natural evolution of an ongoing collaboration** between OBU, IDSIA, SYSTeMS and, more recently, Supélec and DYNAMIXYZ. The first three partners are, it is fair to claim, three of the most visible groups in Europe on Imprecise Probabilities and related theories. Dr Fabio Cuzzolin, Professor Gert deCooman and Dr Antonucci are well recognised researchers in this field, and have a long history of collaboration within the framework of the Society for Imprecise Probabilities (SIPTA), and the organisation of top international conferences in the field such as ISIPTA and ECSQARU.

The Coordinator Dr Fabio Cuzzolin, in particular, is **active in both action/identity recognition and the theory of imprecise probabilities**. He is a recognized leading researcher in the field of belief calculus. He has done extensive work on decision making with imprecise probabilities, and developed an original geometric approach to belief functions and other uncertainty measures. In vision, he is an expert in gait identification, and is supervising Ph.D. students on both part-based, discriminative modelling of actions and the learning of metrics on dynamical models to classify actions dynamics.

In this context, a natural collaboration between OBU and IDSIA has started two years ago circa on the suitability of imprecise graphical models as a way of ensuring better performing, more robust action and gesture recognition algorithm which do take into account dynamics (unlike existing bag of feature methods) and can cope with different sources of uncertainty in terms of occlusions, missing data, insufficient training sets. IDSIA and SYSTeMS are indeed the pioneers of the study of imprecise graphical models which use convex sets of probabilities to cope with such uncertainty.

DYNAMIXYZ and Supélec provide a precious industrial partner which can provide the infrastructure and the expertise to test our research hypotheses in the important VR scenario and kickstart the commercialisation stage. On top of that, they provide to the consortium precious expertise in rendering, capturing and analysing virtual characters and faces. The proposed consortium therefore:

- assembles some of the **best competences in Europe** in imprecise graphical models, model classification, machine learning, action and identity recognition;
- is formed by several partners which have enjoyed a **long standing collaboration** and common networks of relationships;
- has an **overall balance** between theoretical research, application, and industrial development, both at
the level of individual partners and as a whole;
- represents a collection of widely complementary expertises, but at the same time is run by individuals who have multi-disciplinary scientific background in many or all the fields involved.

2.3.3 Suitability and commitment to the tasks assigned

In virtue of its background illustrated above, the Oxford Brookes Vision Group (OBU) is obviously extremely well placed to tackle the tasks assigned to it by the project proposed here:

- identification and classification of dynamical models for action recognition, in particular via metric learning approaches (WP2);
- learning and classification of novel dynamical discriminative models (WP3);
- gathering a corpus of range/stereo videos of actions and gesture in various scenarios (WP4);
- selection and extraction of features from both range cameras and traditional video (WP4);
- application and validation in a variety of different scenarios (WP5).

SYSTeMS and IDSIA are two of the top European groups in Imprecise Probabilities and related theories. Their pioneering work on imprecise (robust) graphical models and the generalisation of the Expectation Maximisation algorithm to convex sets of probabilities is at the current edge of the research in the field. Clearly, it is difficult to think of other research groups in Europe more suitable than SYSTeMS and IDSIA to the crucial tasks assigned to them by the present project:

- imprecise EM and its application to hidden Markov models (WP1);
- imprecise Markov models and general imprecise graphical models (WP1);
- the classification of imprecise graphical models (WP2).

There is in fact a partial overlap between the skills of these two partners, which are both engaged in theoretical research about imprecise probabilities and graphical models. The level of ambition of the project in this area, however (just think of the formulation of a general theory of imprecise graphical models), simply requires the combined effort of two different partners. The ongoing fruitful interaction between the two groups (which have a long tradition of joint works) is certainly going to facilitate a fast and satisfactory achievement of the required tasks.

Supélec and DYNAMIXYZ, besides adding their own specific value to the consortium, form another couple of closely linked partners, with a long record of close and fruitful collaboration in the field of virtual animation and the use and commercialisation of active appearance models in the entertainment industry. As a results, they have won prizes for gesture recognition challenges and established a network of commercial contacts which will be invaluable for this project. They are therefore excellently placed to tackle:

- the capturing of novel testbeds via range cameras for the gaming and animations scenarios (WP4);
- the generation of realistic virtual clones of actors with a specific style (WP5); and
- the recognition of facial expression (WP5).

2.3.4 Complementarity between participants and overall balance

A clear complementarity emerges between the five partners involved in this proposal. IDSIA and SYSTeMS are world leaders on imprecise dynamical models and imprecise probabilities at large. OBU is world leader in computer vision and machine learning, including the target fields of the present proposal: action and gesture recognition, identity recognition from gait. OBU can bring valuable existing infrastructure to the consortium, including range cameras, motion capture devices, computer clusters, and the equipment provided by its Intelligent Transport Systems doctoral programme. The Coordinator is a leading researcher in imprecise probabilities, while both OBU and Supélec/DYNAMIXYZ can provide their industrial expertise and relations, together with the infrastructure necessary for prototype implementation in the manifold scenarios identified by this project. OBU can further support the project via its wealth of industrial connections with world class companies in the field, while IDSIA and SYSTeMS are at the center of the academic community of imprecise probabilities. All partners have extensive networks of collaborations.

As a result, we believe, the consortium is carefully designed and quite well balanced, with respect to the objectives of the project stated in Section 1.1, providing a good mix of expertises and backgrounds, both academic and industrial, in the proportions necessary to ensure the success of the project.
2.4 Resources to be committed

Consortium as a whole.

*Travel to international conferences.* We request funding in order to allow the partners to attend an adequate number of *top international conferences* (such as ICCV, ECCV, NIPS, CVPR, ISIP TA, UAI, AAAI) each year, maximizing in this way the academic dissemination of the results. *Acceptance rates* at these conferences are extremely low (in particular, for vision conferences 15% for poster presentation, around 3-5% for oral presentation) and the *potential impact is enormous*, as papers presented as orals give the authors the chance to talk about their work in front of an audience of 1200 people, among which all the best researchers in the field and many attendees from research divisions of companies such as Microsoft, Mitsubishi, Honeywell, etc, making such a publication of even greater impact than a journal article.

*Coordination activities.*

A sum is requested to fund the foreseen *biennial project meetings*, as long as a number of *visits between the partners* in order to coordinate their actions in a more flexible way. The number of feedbacks and loops between the different parts of the project require a constant coordination, part of which will of necessity have to involve physical travel between the partners. *Junior staff will be encouraged to pay short visits to other partners* in order to gain perspective on the overall project's objectives and targets.

*Dissemination activities.*

We request some money to support our *impact and dissemination plan* (see section 3.2 for a more detailed description). We foresee in particular: customized *seminars/presentations* to be given at companies active in video games/surveillance/biometrics/robotics with the goal of promoting KTP-like industrial partnerships even beyond the lifespan of the project; visits to other research groups in Europe to kickstart *follow-up projects* at both national and European level; meetings with public authorities and *government agencies* in the European Union, focussing on the security applications of our research; the organization of *special sessions* at major conferences to maximize the impact of our breakthroughs; the organisation of *press conferences* to disseminate results to the wider public.

**Partner 1 – OBU**

*Coordinator.* The Coordinator will assign 25% of his time to the project, divided into a 5/9=56% attributed to research activities and 4/9=44% to coordination and management (WP6). More specifically, the Coordinator will directly contribute with his expertise to WPs 2 and 3, and supervise the work of the postdoc and postgraduate research assistants.

*Consortium manager.* A consortium manager will be employed part-time (25% of full time) to assist the Coordinator in managing and coordination consortium activities and relationships with the EC.

*Postdoctoral researcher.* Two postdoctoral researchers are requested full time for the entire lifespan of the project (36 months) to conduct the theoretical research assigned to OBU, i.e., work packages 2 (on the classification of dynamical generative models) and WP3 (on dynamical discriminative modelling).

*Post graduate research assistant.* A postgraduate researcher is requested full time for the entire lifespan of the project (36 months) to conduct the more application-related part of the research assigned to OBU, i.e., work packages 4 (data gathering and feature extraction) and 5 (validation in multiple scenarios). In particular, OBU will work on the gait identification, gaming, robotics and video retrieval scenarios.

*Equipment.* Infrastructure and equipment in the form of *range cameras, Kinect consoles, a motion capture room, an autonomous quad bike, stereo cameras and powerful computer clusters* will be provided by OBU in addition to the EC contribution, greatly contributing to the final success of the project. Software licences for packages such as Matlab and others will be needed.

*Auditing.* We also request an allowance under “subcontracting” for an external audit on the project.
Partner 2 – IDSIA

Unit leader. The unit leader will assign 10% of his time to the project. This resource will be provided by IDSIA, in addition to the EC contribution.

Post Docs. One postdoctoral researcher is requested full time for the entire lifespan of the project (36 months) to conduct the advanced research assigned to IDSIA in the context of WP1 (imprecise generative modelling). Another postdoctoral researcher is requested full time, again for the whole project (36 months) to deal with the theoretical research in work package 2 (dissimilarity measures for imprecise graphical models).

Equipment. Infrastructure and equipment such as computing facilities will be provided by IDSIA in addition to the EC contribution.

Partner 3 – SYSTeMS

Unit leader. The unit leader will assign 7% of his time to the project, divided into a 75% attributed to research activities and 25% to coordination and management. He will contribute to work packages 2 and 3, and supervise the work of the postdocs. This resource will be provided by SYSTeMS, in addition to the EC contribution.

Post Docs. One postdoctoral researcher is requested full time for the entire lifespan of the project (36 months) to conduct part of the theoretically advanced research assigned to SYSTeMS: work packages 1 on imprecise generative modelling and generalized EM. Another postdoctoral researcher is requested full time for the first half of the project (18 months) to deal with the theoretical research in work package 2 (classification of imprecise graphical models).

Equipment. Infrastructure and equipment such as servers and computing (number crunching) facilities will be provided by Ghent University in addition to the EC contribution.

Partner 4 – SUPÉLEC

Unit leader. The unit leader will assign 20% of his time to the project divided into 3/5 for research activities and the 2/5 left to management and coordination. This resource will be provided by Supélec, in addition to the EC contribution.

PhD student. One engineer/Ph.D. student is requested full time at the beginning of the project (24 months) to conduct the research assigned to Supélec in the context of WP4 and WP5, analyzing a real face through multiple RGB-Z cameras.

Equipment. Infrastructure and equipment such as computing facilities and specific RGB-Z cameras will be provided by Supélec in addition to the EC contribution.

Partner 5 – DYNAMIXYZ

Unit leader. The unit leader will assign 10% of his time to the project divided into 50% for research activities and the 50% left to management and coordination. This resource will be provided by Dynamixyz, in addition to the EC contribution.

PhD student. One engineer/Ph.D. student is requested full time for the entire lifespan of the project (36 months) to conduct the research assigned to DYNAMIXYZ in the context of WP4 and WP5, concerning a specific face rigging animatable from the Expressive Space.

Equipment. Infrastructure and equipment such as computing facilities and specific rendering cards will be provided by DYNAMIXYZ in addition to the EC contribution.

As clearly illustrated in Section 1.3 (ii), the requested resources are those we deem necessary to develop both the necessary theoretical advances in the theory of imprecise-probabilistic graphical models and their classification and the development of dynamical discriminative models for action classification and localization, and the more application-directed efforts for acquiring the necessary data, processing it, integrating the results of individual WPs in a single prototype. Impact and dissemination plans are also supported by the above financial plan in a consistent and adequate way.
Section 3. Impact

3.1 Expected impacts listed in the work programme

3.1.1 Connection with expected impacts in the work programme.

The expected impacts listed in the work programme in relation to Objective 2.1: Cognitive Systems and Robotics are the following:

- “Integrated and consolidated scientific foundations for engineering cognitive systems under a variety of physical instantiations.”

Giving its breakthrough nature, and the position of its partners at the top of the respective scientific disciplines, the proposed project is clearly designed to consolidate the scientific foundations within Europe and at worldwide level in a number of key topics of cognitive systems: generative and discriminative learning from labelled training sets, the formulation of novel frameworks for robust action recognition; their testing and application to the variety of physical, societal and industrial scenarios depicted in Section 1.1.

Actions to undertake: this outcome will be a natural consequence of the successful conclusion of the project. An appropriate publication strategy will be crucial in this respect. We will also explore, in the likely case of success, the opportunity of engaging other academic partners in follow-up research projects addressing the further extension and consolidation of the results obtained: for instance, we see scope for more focussed, dedicated projects on manifold learning and discriminative modelling, and wider (possibly IP) projects on the integration between vision and robotics (in terms of both terrestrial and aerial vehicles), building on our existing record in autonomous navigation.

- “Significant increase in the quality of service of such systems and of their sustainability in terms of, for instance, energy consumption, usability and serviceability, through the integration of cognitive capabilities.”

An intelligent use of the proposed techniques has the potential to contribute to energy saving and environmental protection. In surveillance, for instance, automatic event recognition could lead to more efficient security approaches in which humans do have to waste their time for long hours in front of static scenes, but are only alerted when some unusual activity is actually taking place. Efficient tools for gesture recognition can help much in supporting disabled or elderly people, much improving their quality of life. In video data retrieval, manual search could be significantly sped up using automatic retrieval techniques, resulting in both energy saving and increased productivity in the high-tech commercial sectors involved. As a general remark, efficient automatic or semi-automatic (with human assistance) action recognition has enormous potential for improving the competitiveness of European players in all involved scenarios, via significant gains in productivity.

- “Innovation capacity in a wide range of application domains through the integration of cognitive capabilities.”

The huge number of scenarios in which the need for robust action and activity recognition arise have been widely discussed when introducing and motivating the problem: they are reflected in the tasks the central work package 5 is articulated into: gait identification, entertainment and interactive video games, video retrieval over the internet, autonomous navigation, animation and facial expression analysis.

In all these scenarios (and in many others which we mentioned but could not include here for sheer lack of time and resources, think for instance of human computer interaction, smart homes, and semi-automatic care for the elderly) a successful outcome of this project will contribute enormously in terms of both cutting-edge innovation and economic boost for the economic and societal sectors involved, all the more crucial in times of economic uncertainty and recession.

Companies like Google are heavily investing on video data mining; Street View is just an example...
of a commercial application in which autonomous vehicles may need to interact with pedestrians; behavioural biometrics is at the forefront of security and surveillance (it was indeed initiated by the US agency DARPA); many other examples can be given to show how many sectors this project could impact if successful.

- “Improved competitive position of the robotics industry in existing and emerging markets for instance in the following sectors: manufacturing; professional and domestic services; assistance and co-working, production, logistics and transport, construction, maintenance and repair, search and rescue, exploration and inspection, systems monitoring and control, consumer robotics, education and entertainment.”

We have been long arguing how the market potential of action and gesture recognition applications to surveillance and activity recognition, behavioural biometrics, human-computer interfaces and virtual reality, video retrieval over the internet is just too big to be described extensively here. The case of motion-based video games interfaces is enlightening. Microsoft's Kinect console, with its controller-free gaming experience is already revolutionizing the whole market of interactive video games and consoles (see http://www.xbox.com/en-us/live/projectnatal/ for some amazing demos). In the entertainment industry, intelligent action recognition could significantly improve user experience, “spicing up” games which currently limit themselves to tracking the user's. The same ideas could be applied to educational games for young children and adults alike; educations tools could be developed along the very same lines.

Assistance to the elderly, though not explicitly covered by WP5, may be kick-started in the course of the project in the form of a spin-off project or transfer partnership.

Novel biometric techniques based on gait are likely to gain market share in the current security scenarios. Most biometric companies focus at the moment on cooperative techniques such as face or iris recognition: investing in behavioural, non-cooperative biometrics before the rest of the market could provide them with a significant competitive advantage. “In both the identity management and security arenas, the use of biometric technology is increasing apace ... the world biometrics market has expanded exponentially. Annual growth is forecast at 33% between the years 2000 and 2010. Europe is expected to have the fastest growing biometrics market by 2010 ... The Intellect Association for Biometrics (IAfB) is the UK body that represents companies developing these technologies ... has fostered close ties with the UK Border Agency and Home Office.” (Biometrics, November 2008).

Companies such as Google are investing huge money on image and video browsing as the next level in the browsing experience. A successful outcome of this project could have impact and consequences on the competitiveness of European industries on all these fast-growing markets.

The entertainment and advertising industry is interested in cloning actors, e.g. for Digital SFX movies such as “Avatar” or “The Curious case of Benjamin Button”. The market is huge: “Beyond cinema and TV actors, Digital Doubles are bound to reach all types of celebrities - sportsmen, singers, fashion models, TV and business people… - over the different media – advertising, video clips, video games cinematics… - before their natural extension to interactive and real-time applications – video-games, broadcast TV shows, web applications…” [136].

Actions to undertake: the degree to which the project will enjoy such a commercial impact will be the result of a close engagement with world class companies and our own industrial partner. Resources and exploitation actions have been explicitly requested and put in place to this purpose (see Section 3.2 for a very detailed description).

- “Consensus by industry on the need (or not) for particular standards. More widely accepted benchmarks. Strengthened links between industry and academia.”

The proposed project brings together top R&D groups, academic institutions and award-winning industry partners such as DYNAMIXYZ. In addition, the partners have a wealth of industrial contacts that will further contribute to a more close interaction between industry and academia. For instance, OBU's Computer Vision group has an established record for exploiting IP and interacting with companies such as Microsoft Research Europe, Sony Entertainments Europe, Vicon, 2d3, HMGCC, Yotta and Sharp. The group leader Philip Torr was involved in setting up the computer
vision group at Microsoft Research Cambridge UK. He was also involved in the creation of the start-up company 2d3 (http://www.2d3.com/), part of the Oxford Metrics Group (OMG), whose first product, “boujou”, is still used by special effects studios all over the world and has been used on the special effects of almost every major feature film in the last five years, including the “Harry Potter” and “Lord of the Rings” series. Yotta (http://www.yottadcl.com/) is another OMG company currently funding Ph.D. studentships at the Oxford Brookes vision group. In the course of its cooperation with Armasuisse (the R&D agency of the Swiss army) IDSIA developed a number of intelligent systems currently used by the Swiss air force to enforce no-fly zones as well as automatic reasoning tools for social networking in the music business, developed by a Swiss start-up company (www.stagend.com) Many other examples of successful industry collaborations and exploitation agreements could be cited.

The new datasets acquired in the course of the project will likely have great impact as new benchmarks for the whole community.

Actions to undertake: the consortium will exploit its already considerable academic and industrial links to maximise the dissemination of the results of the project in terms of theoretical advances, public domain code, gathered datasets. We will both build on the existing network by engaging companies on a personal level, and work on building new strategic contacts via participation to dedicated events organised by the EU and other bodies. We envisage likely exploitations of our results on both the national and European level, for instance (in the UK) by setting up Knowledge Transfer Partnerships (KTPs) with companies expressing an interest.

Again, many more details on the measures we plan to enact to maximise the expected impacts can be found in Section 3.2.

3.1.2 External factors affecting the potential impacts

In principle, we see three possible sources of threat to the concrete realization of these impacts: the lack of empirical support to the scientific hypothesis at the foundation of this project; the lack of data on which to extensively validate the results in order to bring them close to real-world deployment; an unfriendly societal and industrial response to the research conducted here.

None of these constitute, in our view, significant problems.

Scientific hypothesis.

Obviously, the potential impacts listed above can be reached only in the case of a successful termination of the project. The overall, underlying assumption is that encoding dynamics is important to achieve robust action recognition in both the discriminative and the generative approach to the problem, and that in the latter case imprecise-probabilistic graphical models are suitable to cope with the sources of nuisance, the inherent variability and the overfitting due to the limited size of the training sets that characterise the problem. Given our already considerable and growing expertise in all these matters, however, evidence does support this key assumption.

As argued in Section 1, on new datasets with 51 action categories bag-of-features models achieve classification results of just above 20%, suggesting that these approaches do not fully represent the complexities of human actions. Kuehne al. have shown that this drop in performance with respect to previous datasets is most likely due to the increased number of action categories. The most recent literature [104] has shown that, while in the UCF Sports dataset [109] human static joint locations alone are sufficient to classify sports actions with an accuracy of over 98%, the same experiment on HMDB51 performs quite poorly (35%). This indicates that it may be necessary to take more into account action dynamics, in accordance with psychophysical experiments in which both motion and shape are critical for visual recognition of human actions [110, 111].

Availability of data.

The chance that suitable data are not available to thoroughly validate our methodologies is neutralised by both the availability of a number of datasets of significant size and complexity (though limited to some extent as discuss for WP4), and that of proprietary equipment which will allow us to acquire testbeds in multiple modalities, with a focus on all the aspects we feel current datasets still neglect: complex activities, range data, large number of action categories, etcetera.
Societal and industrial context.

Negative external factors should not be able to influence the success of the project either. As we argued throughout this proposal, current market conditions, societal and technological developments are very favourable to the application of the novel methodologies we propose to develop here to the various scenarios depicted in WP5 and further discussed above. The enormous interest by companies, governments and ordinary European citizens for natural human computer interaction, improved security levels, novel biometric techniques, and more “fun” entertainment systems will ensure that a positive scientific outcome of the project is guaranteed a rapid diffusion and, possibly, commercialised via focussed industrial partnerships.

3.1.3 Why a European approach

To maximise impact and effectiveness of a research project, the best talents and skills have to be recruited. In the case of our project, we identified groups which are at the top of their respective fields: OBU vision group in Computer Vision and Machine Learning, IDSIA Imprecise Probabilities group, Ghent's SYSTeMS group in imprecise probabilities and inference in robust statistical models, Supélec and Dynamixyz in virtual animations and face expression recognition. The partners happen to belong to different countries: the United Kingdom, Belgium, Switzerland and France, but the composition of the consortium is the direct consequence of the objectives of the project: this illustrates how the most ambitious feats are often only possible at European level, where the necessary competences can be found.

The reasons that have led to a consortium of this specific scale and composition have been discussed in much detail in Section 2.3, but we think useful to make our case further here, in terms of integration of both skills and existing equipment which, again, could only take place at European level.

Integration of skills and related projects.

The project proposed here is related to other projects proposed or underway by the partners at national level. OBU School of Technology, for instance, is currently funding a PhD studentship in Computer Vision and Robotics/Autonomous Navigation under the new Intelligent Transport Systems Doctoral Training Programme. The research involved requires solving a number of sophisticated vision problems, such as the recognition of static (signs, cars) and moving (people, animals, bicycles) obstacles along the vehicle’s path or in its vicinity, which are quite related to the current project. In addition, the Coordinator's currently running EPSRC First Grant proposal is concerned with identity recognition from gait via tensorial methods, as the latter are able to describe the various covariate factors which affect recognition as a linear mixture of independent dimensions.

As for IDSIA, the institute has received support from the Swiss National Foundation for a Ph.D. and two post-docs involved in research on imprecise probability. Dr Antonucci, the reference person for the project at IDSIA has received a 4-years grant from the Swiss Army for the application of credal networks to military security problems. He is also supervising a Master student working on the application of (precise) HMMs to action recognition. With respect to topics related to this project, SYSTeMS has received support for two research projects at the national level which deal with imprecise-probabilistic dynamics, and probabilistic graphical networks in bioinformatics. It is also currently a partner in a Multidisciplinary Research Partnership called N2N (from Nucleotides to Networks), a new initiative to spearhead research on Bioinformatics and Systems Biology at Ghent University.

Supélec participates to a national project (Replica) in which its task is to analyse the expressions of young people with cerebral palsy. A Supélec Ph.D. studentship is funded by this project. The objective is to analyse the dynamics of the expressions on such people all along their rehabilitation process, in order to evaluate it in an objective manner. This work is obviously highly related to that proposed here in dynamic modelling. Another Supélec Ph.D. student is funded by the Brittany Region of France through the project PME Sketcher (2012-2014), in which the expressions of a teacher are analysed in real time via a RGB-Z camera. Faces are modelled in 3D and depth information is integrated within the framework of Active Appearance Models. This work will be exploited in Dynact, since we plan to analyse real faces with the help of several RGB-Z cameras in order to optimize a 3D Active Appearance Model.

Integration of equipment.

In addition to gathering the necessary specialized skills to achieve the objectives of the project, this consortium also provides through its partners a wealth of equipment which is not only valuable but crucially necessary for the project. Range cameras are already in possession of OBU and DYN, commercial products for rendering of realistic animations are provided by DYN, powerful computer clusters are available at OBU.
and SYSTEms, an autonomous quod bike is being assembled at OBU under the Intelligent Transport Systems doctoral programme. As far as software is concerned, a vast wealth of routines for feature extraction and selection, image segmentation, clustering, model identification is brought by OBU as a dowry to the project. The same can be said for imprecise modeling at IDSIA and SYSTEms, and for the tracking of active appearance models and the definition of expressions spaces at Supélec and Dynamixyz.

In conclusion, the individual partners are already successfully engaged in partial aspects of the composite research theme proposed here. So far, however, they could not aim at proposing a comprehensive solution due to lack of necessary skills and workforce. OBU does not possess the expertise on imprecise-probabilistic graphical models; IDSIA and SYSTEms have no background on discriminative models, or feature extraction. Supélec and Dynamixyz provide competences and links in virtual reality, OBU in gaming, robotics and biometrics. Brought together, they can shoot for a real breakthrough in action modeling and recognition.

3.2 Dissemination and/or exploitation of project results, and management of intellectual property

3.2.1 Beneficiaries of the proposed project

The research project proposed here has a number of clear beneficiaries, which will be targeted to maximise impact and dissemination of its outcomes, both academic and industrial.

Academic beneficiaries.
As the proposed research spans both theory and application, we foresee not only a clear potential impact on the ever-expanding community of human action, gesture and activity recognition, but also a methodological impact in a number of other fields of applied science in which time series and manifold learning play an important role.

Impact on action recognition.
While dynamical models have been widely employed in action, gesture and gait recognition, their potential has been partially impaired by a rather naive approach to classification. The framework we propose aims at exploiting the full abstraction power of complex dynamical graphical models, both precise and imprecise, by applying metric learning techniques and sophisticated classification techniques in order to achieve greater robustness and capability of overcoming the many critical issues in action recognition. As such, this study could lead to much subsequent research in the field along these lines.

The same holds for the novel, structured discriminative models we propose and for which we have great expectations, for we believe they could really revolutionise the field.

Impact on imprecise probability and artificial intelligence.
The development of a comprehensive theory of imprecise probabilistic graphical models, in particular iHMMs and imprecise MRFs will be a truly groundbreaking contribution to the field of imprecise probability, greatly contributing to its adoption by a growing number of practitioners as an alternative to more traditional Bayesian methods.

Impact on manifold learning.
The proposed manifold learning scheme constitutes, if successful, a decisive step towards the extension of distance learning techniques to the case of data living in complex, nonlinear spaces. In the last ten years numerous studies have been conducted on the question of learning the “best” distance function for a given classification problem, in which however a linear mapping is optimized upon. This study directly contributes to and to same extent pioneers the extension of this approach to the nonlinear case, and its application to time series in the form of dynamical models. The application of cutting edge compressed sensing techniques and their adaptation to complex objects such as generative models is also extremely interesting and promising.

Industry and high tech companies.
Within the private sector companies whose core business is in human-machine interface, semi-automatic surveillance and biometrics, entertainment and virtual reality, and data retrieval would be the primary beneficiaries. Most biometrics companies, for instance, focus at the moment on cooperative biometrics such as face or iris recognition: investing in behavioural techniques ahead of the rest of the market could provide them with a significant competitive edge. The game and entertainment industry could be shaken up by a new generation of games which allow for real gestural interaction between user and game characters. Indeed, as
already explained, the consortium already has a long history of collaborations and partnerships with such companies. The elderly care industry could be revolutionized by the diffusion of intelligent home technologies, of which activity recognition is a crucial part. The same can be said for the all the other scenarios of WP5.

Being based on graphical models, the algorithm developed for efficient generative recognition of videos can naturally support a distributed architecture: this is very useful when moving from software to firmware implementation, a particularly important aspect when it comes to potential industrial partnerships. The competitiveness of all these actors could potentially be hugely boosted by the adoption of cutting edge technologies developed in this project, to the economic competitive advantage of the Union.

**National and European policy-makers.**

We also expect policy-makers and government agencies to be attracted by the idea of novel surveillance and biometric systems able to improve the general level of security in the continent, especially in the current, uncertain times, and that of sensitive areas in particular. Examples are airport management authorities such as BAA, railway companies, underground and public transport authorities (e.g., Transport for London). The infrastructure is basically already there, as millions of active CCTV cameras are active throughout the European Union and the European Economic Space. A research partner of IDSIA, Armasuisse, the research department of the Swiss army, has already expressed potential interest in the outcome of the project. Similar considerations can be made about national armies and intergovernmental military alliances, very interested in the general topic of situational awareness, which would clearly benefit from a successful outcome of the proposed project.

**The European citizen.**

The wider European public will of course be the ultimate beneficiary of any improvement in the security level of public places and transport. The EU citizen’s quality of life will arguably benefit too, in the longer term, by the deployment of behavioural biometric systems as anti-terrorist measures. In addition, robust action recognition will likely revolutionise many aspects of ordinary citizens’ life, in terms of the way they spend their spare time (e.g., interactive gaming), they go through their daily tasks at work (e.g., new human-computer interfaces), and live in their homes (as in the smart room scenarios).

### 3.2.2 Measures proposed for the dissemination/exploitation of the results

In order to reach the various groups of beneficiaries identified above, we plan a coordinated set of actions with the purpose of engaging them in a variety of ways.

**Figure 13.** Impact actions envisaged by Dynact: special issues, workshops, grand challenges (some images from the Pascal VOC challenge), contact with industrial groups (Google Santa Monica), dedicated web site.

**Publication strategy.** The obvious step in a research project is the publication of its scientific results in international journals and top conferences. This is all the more obvious for a high-profile project such as the proposed one, which involves a number of scientific communities active on imprecise probabilities, computer vision, image processing, gait identification, machine learning, virtual reality.

Accordingly, target journals include all the top venues in computer vision, machine learning and artificial intelligence:

- IEEE Transactions on Pattern Recognition and Machine Intelligence (I.F. 4.38);
- The International Journal of Computer Vision (I.F. 3.508);
- Journal of Computer Vision and Image Understanding;
Artificial Intelligence (I.F. 3.035);
- the IEEE Transactions on Systems, Man, and Cybernetics Parts A,B,C (I.F. 3.01, 2.02)
- the International Journal of Approximate Reasoning (I.F. 2.09);
- the IEEE Transactions on Fuzzy Systems (I.F. 3.34);

Notice that vision journals are the top venues in absolute terms in Computer Science, further boosting the academic impact of this project.

Target international conferences include the following top computer vision conferences:

- the IEEE International Conference on Computer Vision (ICCV);
- the IEEE Conference on Computer Vision and Pattern Recognition (CVPR);
- the IEEE European Conference on Computer Vision (ECCV);
- the British Machine Vision Conference (BMVC).

Oxford Brookes is deeply involved in the organisation of these venues. OBU is part of the steering committee of the upcoming ICCV'13 event in Sidney, Australia, through his founder Philip Torr. The Oxford Brookes vision group has also organized BMVC in recent years, and its members are regular reviewers for all such conferences.

The top conferences in Artificial Intelligence are considered to be:

- Uncertainty in Artificial Intelligence (UAI);
- the Conference of the National Association for Artificial Intelligence (AAAI);
- the International Joint Conference on Artificial Intelligence (IJCAI).

Dr Antonucci is a permanent member of the Programme Committee for UAI, which is chaired by Professor Fabio Cozman, a professional partners of OBU, IDSIA and SYSTeMs. He has also been involved in tutorials at AAAI. As for the top machine learning venues, we target:

- Neural Information and Processing Systems (NIPS); and
- International Conference on Machine Learning (ICML).

OBU's record at NIPS is outstanding, having recently won a Honorable Mention at NIPS'08.

Finally, the most important symposiums in Imprecise Probabilities are:

- the International Symposium on Imprecise Probabilities and Their Applications (ISIPTA);
- the European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU);
- the International Conference on the Theory of Belief Functions (BELIEF).

OBU's Fabio Cuzzolin, IDSIA's Antonucci and SYSTeMs' deCooman are permanently members of the programme committees of all these conferences.

As it is to be expected, conference papers will be targeted in the first year(s) of the project, to later lead to journal publications. Two of the three major vision venues run every single year, as UAI, AAAI, NIPS and ICML do. Imprecise probability conferences are normally organized every other year.

Organisation of special sessions and workshops at selected international conferences. Many major computer vision conferences such as ICCV, ECCV and CVPR regularly organise one or two-day workshops on selected topics. Relevant examples are the Action Recognition Challenges CVPR Workshop (http://www.umiacs.umd.edu/conferences/cvpr2011/ARC/) and the analogous SMC Workshop (http://www.opportunity-project.eu/workshopSMC).

We intend to organise such workshop sessions on the project (possibly in collaboration and accordance with similar initiatives) in order to further disseminate our results within the academic community. Major such
workshops normally are attended by some 50 specialists, and lead to the publication of books and collections with the potential of further disseminate the results of the project. These occasions will provide natural venue in which to establish new academic links and prepare the ground for spin-off or follow-up projects and national and European level.

Relevant expertise is present in our consortium: Dr Antonucci has organised two special sessions on imprecise probabilistic graphical models at the IPMU conference (International Conference on Information Processing and Managing of Uncertainty in Knowledge-based Systems) in both 2010 and 2012.

**Organization of Grand Challenges on action recognition in the different scenarios.** In recent years, the PASCAL VOC grand challenge ([http://pascallin.ecs.soton.ac.uk/challenges/VOC/](http://pascallin.ecs.soton.ac.uk/challenges/VOC/)) on visual object classes recognition has greatly contributed to revolutionise computer vision. The VOC challenge was kick-started by a collaborative project such as the one proposed here. In recent times, an “Action Classification taster” has been introduced which copes with recognition from single images.

A similar grand challenge on full action recognition would be a natural development along this line. For its characteristic, this project has a clear potential to given a similar contribution to action recognition in terms of both methodologies and test-beds: the launch of a grand challenge on this theme could potently magnify its impact.

**Organisation of special issues of international journals** on the project's topic. Thanks to the role of many of the partners in the Editorial Boards of several international journals, we plan to organise special issues on the various topics covered by the project in the course of the years. In particular, an issue on imprecise-probabilistic graphical models could fit very well IJAR or AIJ; spatio-temporal discriminative models could be the topic of a PAMI or Machine Learning special issue; an action or activity recognition challenge could result in a special issue of IJCV or PAMI.

**Building on existing nexwork of links.** We also plan to build on the consortium's already strong network of existing professional links in order to maximise the impact of this project. For instance, OBU is currently engaged with organisations interested in security matters, such as HMGCC ([http://www.hmgcc.gov.uk/](http://www.hmgcc.gov.uk/)), the UK government centre of excellence whose aim is to design and develop secure communication systems, hardware and software specifically for HM Government use, both at home and overseas. HMGCC could be involved towards the final part of the project in order to test government interest in novel biometrics. Concerning action recognition, OBU can bring added value to the project given its extensive links with world class companies active in the area such as Microsoft Research Europe, Google, Yotta, Sony Entertainment. It will be quite straightforward to attract the interest of such industrial behemoths in case the performances of our prototype meet our expectations. Contacts will be made informally (at least at first) to probe potential interest. Sony and Oxford Metric Group have already manifested interest in writing support letters for UK level EPSRC project proposals, are since long time collaborating with OBU on the supervision of several Ph.D. students, and have a history of providing equipment in the form of traditional and range cameras.

**Building new industrial links.** At the same time, new industrial links with companies active in recognition will have to be pursued in order to establish Knowledge Transfer Protocol agreements, proofs of concept or patents.

An example is a tentative collaboration brokered by Brookes’ consultant Eugene Sweeney with DECTEL Inc. ([http://www.dectel.co.uk/](http://www.dectel.co.uk/)), a company active in video surveillance. Many other companies active in biometrics, vision applications and gait analysis are present in the European Union. They will be targeted via “customised” seminars in order to present the results obtained by the project in the most interesting way from the specific point of view and focus of each individual company. Microsoft Research in the person of Andrew Fitzgibbon, involved in the Kinect project, is being contacted to verify their interest in a formal collaboration. As commercial enterprises are mostly focused on simpler, indoor environment (as in the gaming scenario) setting new industrial partnerships could already be possible during the second part of the project. The Coordinator plans to exploit his personal contacts at Google Zurich (M. Marszalek) and Santa Monica (A. Bissacco) in this sense. In addition, he plans to attend the next ICT conference in Brussels as an opportunity for creating new links at EU level, to be later exploited for commercialisation or patenting.

**Engaging government bodies and security agencies.** Given its repercussions in security applications such as semi-automatic event detection, or identity recognition from gait, we believe a range of government
bodies, security agencies (HMGCC is just an example), and utility/transportation providers (such as BAA in the UK) will be interested in testing the effectiveness of our techniques in improving the security levels of their countries or facilities.

Given the more challenging nature of surveillance applications which are the focus of the interest of public bodies and authorities, we plan to engage them towards the very end of the project, in case the results strongly support the feasibility of an outdoor deployment of activity recognition systems or gait-based identification prototypes. As a first step, national governments could be invited to sponsor a prototype to be thoroughly tested on the field. We will arrange preliminary meetings with these public or semi-public entities in the last six months of the project, in which we will display the performance and robustness of our prototype.

Engaging intermediate regional or national bodies. Actions will also be taken through intermediate bodies such as, for instance, the Enterprise Europe Network (www_enterprise_europe swirl.org.uk), an organisation promoting contact opportunities for business partners and supporting intellectual property management and access to European funding. The Coordinator has indeed recently attended an ICT info day at HP Labs in Bristol in April 2010, and established there links with potential partners, which could be useful to kickstart follow up projects. Another organisation we plan to actively involve is the Knowledge Transfer Network (https://ktn innovate uk.org/), a platform set up by the UK Technology Strategy Board which “provides an effective and powerful way for you to collaborate online, network and share knowledge with other innovators”.

Exploiting the partners' academic links to kickstart follow-up projects at national or European level. STREPs, by definition, are small scale projects designed to focus on specific, well defined problems. Clearly, though, the topics addressed by this project, both on the theoretical side (the formalisation and consolidation of inference techniques for dynamical models which handle uncertainty via sets of probability distributions, and the formulation of discriminative models which incorporate dynamics) and on the side of applications (action, gesture, activity, or identity recognition from motion) are at the cutting edge of current research in the area, and their study will necessarily proceed far beyond the lifespan of any project. Indeed, the Coordinator is seeking UK EPSRC funding on discriminative modelling and manifold learning for dynamical models on a separate three-year project [16], and has been awarded funds by EPSRC for a two year project on gait identification. It makes therefore sense to plan, towards the end of the project, a number of subsequent initiatives at both national and, possibly, European level.

The existing consortium enjoys a large network of contacts throughout the European continent, including: INRIA (France), Universitat Pompeu Fabra (Spain), EPFL and IDIAP (Switzerland), Politecnico di Milano (Italy) and others, that can and will be exploited in this sense. For instance, Barcelona’s CILAB (http://cilab2.upf.edu/index2.html), a partner of OBU in a separate upcoming FP7 FET proposal on large scale manifold learning and shape analysis, is active in blood vessel and heart simulation via bilinear modelling, and enjoys intensive collaborations with several hospitals. Oxford University's professor Thomas Lukasiewicz is working with the Coordinator on a Google Research Award on video retrieval under uncertainty that could lead to further research on the automatic description of videos for browsing purposes. A three-year national French project FUI-11 OSEO / DG CIs on cloning has started in January 2012. We have no doubts that the company involved in this project (Agence de Doublure Numérique) will be really interested by the results of Dynact.

Web site. Besides seminars targeting audiences such as companies and other organisations, we attribute great importance to the dissemination of our results through a web site dedicated to the project, which will be designed in order to attract not only academic interest, but also knowledge transfer partners and will be publicised accordingly. The web site will serve several purposes:

- distribute the data and algorithms collected/generated by each individual partner to all the others, through a secure server;
- disseminate the scientific results to the wider academic community (with a focus on the communities of imprecise probabilities, machine learning, action and gesture recognition, and gait identification) using a different channel than that of publication on journals and conferences: this could also improve the scientific impact in terms of (Google Scholar) citations;
- make the final libraries of algorithms for imprecise model identification and classification, classification via dynamic discriminative models, and feature extraction publicly available, again to
maximise the impact of the project: the creation of new baseline algorithms could considerably enhance the visibility of the latter;

- make the datasets gathered in the course of the project publicly available as new benchmarks to the academic community, in this way boosting the visibility of the whole enterprise; these datasets will be either downloadable via FTP server or by direct shipping via external USB hard drive;
- launch grand challenges on the various scenarios which will be investigated throughout the project;
- attract new industrial and academic partners with the goal of setting up patents or industrial partnerships;
- attract new academic partners to kick-start follow-up projects at national or European level.

3.2.3 Timescales

Realistic timescales for the benefits discussed in Section 3.2.1 will vary with the specific target. In terms of academic impact, a first round of conference publications should be ready at the end of the first year, while journal articles will likely take two or three years to reach publication stage. The use of vehicles such as special issues of international journals linked to consortium partners could help to speed up the process.

Industrial involvement is built-in the proposed project, thank to the presence of DYNAMIXYZ with its infrastructure and network of industrial relations. However, other partners will be sought to further increase impact as explained exhaustively above.

Existing links will take no time to be activated when the consortium deems appropriate to do so. As previously argued, new industrial links are already being explored; this activity will continue throughout the project. The same can be said of intermediate entities and organisations.

To start seeing commercial applications to biometrics or surveillance proper, initially in a limited context such as for instance biometric access, an additional two to three years of R&D after the end of the proposed research might be needed, possibly in partnership with a company specialised in this area. It is safer to assume a longer time scale when targeting a wider application to security in public areas: government or agency support will in this case be crucial.

The timescale for benefits to the wider public basically depends on how quickly the developed techniques can be integrated in actual entertainment products (consoles) or search engines. Within the gaming scenario, OBU's existing industrial partners such as Sony could much contribute to speeding up the dissemination of the results and the necessary technology transfer.

As for video retrieval, we envisage a potentially much swifter dissemination process due to fact that no expensive equipment of sort is needed. Furthermore, Dr Cuzzolin's personal contacts at Google could be put to good use to test our algorithms on the field.

The engagement of other academic partners at national or European level will be kick-started in the second half of the project, in order to be ready for the preparation of follow-up proposals before the end of the current one. However, as already pointed out, an extensive network of contacts is already at work and could be activated in no time.

3.2.4 Intellectual Property management

Intellectual Property Rights management and exploitation will be managed (in close contact with the Coordinator and the Administrative Manager) by the Research and Business Development Office (RBDO) at Oxford Brookes University (OBU), which has access to financial and other resources to enable Intellectual Property and its commercial exploitation to be effectively managed, whilst maximising the widespread dissemination of the research results. This includes, as appropriate, finance for patenting and proof of concept funding; Intellectual Property, technology and market assessment; resources for defining and implementing a commercialisation strategy though licensing, start-up company or other routes.

RBDO has a strong track record of commercialisation of its intellectual property. Income from licences was £1.5M in 2011 and this ranks the University in the top 10 in the UK of all universities for royalty income. OBU through RBDO holds a total portfolio of 20 patents. The University also supports the creation of spin out companies when appropriate and has some successful examples.

If the outcome of the project in terms of implementation (see workpackage 5) will meet our expectations, we will consider the possibility of patenting the associated systems and algorithms.
Section 4. Ethical Issues
(No maximum length for Section 4 – depends on the number and complexity of the ethical issues involved)

While the proposed research is about the recognition of people’s behaviour, it does not involve the tracking of individuals in any way, therefore it does not pose any issue of privacy violations.

Notes:
1. For further information on ethical issues relevant to ICT, see annex 5 of the Guide for applicants.
2. Only in exceptional cases will additional information be sought for clarification, which means that any ethical review will be performed solely on the basis of the information available in your proposal.
<table>
<thead>
<tr>
<th>Ethical Issues</th>
<th>YES</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Informed Consent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve children?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve patients or persons not able to give consent?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve adult healthy volunteers?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve Human Genetic Material?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve Human biological samples?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve Human data collection?</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Research on Human embryo/foetus</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve Human Embryos?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve Human Foetal Tissue / Cells?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve Human Embryonic Stem Cells?</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Privacy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve processing of genetic information or personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve tracking the location or observation of people?</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Research on Animals</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve research on animals?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Are those animals transgenic small laboratory animals?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Are those animals transgenic farm animals?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Are those animals cloned farm animals?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Are those animals non-human primates?</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Research Involving Developing Countries</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Use of local resources (genetic, animal, plant etc)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Impact on local community</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dual Use</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Research having direct military application</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Research having the potential for terrorist abuse</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ICT Implants</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Does the proposal involve clinical trials of ICT implants?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL

| | X |

Proposal Part B: page [62] of [67]
Annex I: References

[29] D. Han, L. Bo, and C. Sminchisescu, Selection and context for action recognition, ICCV’09.


