European Cooperation in the field of Scientific and Technical Research

- COST -

Brussels, 22 November 2013

COST 067/13

MEMORANDUM OF UNDERSTANDING

Subject :	Memorandum of Understanding for the implementation of a European Concerted Research Action designated as COST Action IC1307: The European Network on Integrating Vision and Language (iV&L Net)

Delegations will find attached the Memorandum of Understanding for COST Action IC1307 as approved by the COST Committee of Senior Officials (CSO) at its 188th meeting on 14 November 2013.

MEMORANDUM OF UNDERSTANDING
For the implementation of a European Concerted Research Action designated as

COST Action IC1307
THE EUROPEAN NETWORK ON INTEGRATING VISION AND LANGUAGE
(IV&L NET)

The Parties to this Memorandum of Understanding, declaring their common intention to participate in the concerted Action referred to above and described in the technical Annex to the Memorandum, have reached the following understanding:

1. The Action will be carried out in accordance with the provisions of document COST 4114/13 "COST Action Management" and document COST 4112/13 "Rules for Participation in and Implementation of COST Activities" , or in any new document amending or replacing them, the contents of which the Parties are fully aware of.

2. The main objective of the Action is to create an interdisciplinary European vision-language research network targeting scientific advances and societal benefits in four focus themes addressing Integrated Modelling of Vision and Language, and Applications of Integrated Models including Image/Video Description and Search.

3. The economic dimension of the activities carried out under the Action has been estimated, on the basis of information available during the planning of the Action, at EUR 36 million in 2013 prices.

4. The Memorandum of Understanding will take effect on being accepted by at least five Parties.

5. The Memorandum of Understanding will remain in force for a period of 4 years, calculated from the date of the first meeting of the Management Committee, unless the duration of the Action is modified according to the provisions of section *2. Changes to a COST Action* in the document COST 4114/13.

A. ABSTRACT AND KEYWORDS

The explosive growth of visual and textual data (both on the World Wide Web and held in private repositories by diverse institutions and companies) has led to urgent requirements in terms of search, processing and management of digital content. Solutions for providing access to or mining such data depend on the semantic gap between vision and language being bridged, which in turn calls for expertise from two so far unconnected fields: Computer Vision (CV) and Natural Language Processing (NLP). The central goal of iV&L Net is to build a European CV/NLP research community, targeting 4 focus themes: (i) Integrated Modelling of Vision and Language for CV and NLP Tasks; (ii) Applications of Integrated Models; (iii) Automatic Generation of Image & Video Descriptions; and (iv) Semantic Image & Video Search. iV&L Net will organise annual conferences, technical meetings, partner visits, data/task benchmarking, and industry/end-user liaison. Europe has many of the world's leading CV and NLP researchers. Tapping into this expertise, and bringing the collaboration, networking and community building enabled by COST Actions to bear, iV&L Net will have substantial impact, in terms of advances in both theory/methodology and real world technologies

Keywords: Computer Vision and Language Processing, Integrated Modelling of Vision and Language, Automatic Annotation of Image/Video, Language-driven Image/Video Search, Cross-media Data Management and Mining.

B. BACKGROUND

B.1 General background

The amount of digital information available on the Web and stored in diverse types of data repositories is growing at an ever faster pace. Digital information has always meant text, but in the 21st century, it increasingly means visual content as capturing digital visual content has become very cheap and easy. This development has been rapid and has resulted in a situation where computational solutions are lagging behind a diverse range of pressing needs relating to the search, processing, accessibility, mining and management of visual content. Thus, institutions such as hospitals and police forces are unable to effectively utilise the massive amounts of images and video footage they produce daily: for example, more than 35,000 images are produced by the Radiology Department of a single Geneva hospital group per day, creating enormous problems in terms of data processing and search apart from other issues such as data storage and management.

There are more than 4.2 million CCTV cameras in the UK, but the police does not have the capacity to use the footage generated except for the most serious of crimes. Visual content also means reduced accessibility: European disability discrimination legislation and World Wide Web Consortium (W3C) guidelines are clear that visually impaired internet users should have equal access to websites, but in reality accessibility is getting worse; e.g. 75% of FTSE 100 company websites do not meet even the minimum W3C accessibility requirements, and website owners are being taken to court over accessibility.

Progress in these areas is hampered by the big, and as yet un-bridged, semantic gap between visual content and meaning. Recognizing (and consequently correctly indexing) content in visual data is currently limited to the recognition of people, objects, buildings or scenes for which manually annotated training data is available and even then the training data might be insufficient to capture all variations and ambiguous renderings of the target objects. Advanced solutions for image/video search, retrieval, automatic description and similar challenges require the semantic gap to be bridged, and this in turn calls for expertise from two currently disconnected research fields: computer vision (CV) and natural language processing (NLP).

Recently, the benefits that result when expertise from these two areas is combined are beginning to become apparent (see Section B.2 for details). The current situation is that pockets of research are emerging in different European countries investigating the potential benefits of jointly modelling aspects of language and image/video processing, and bringing techniques from NLP to bear on CV tasks and vice versa. What is needed now is a forum for these researchers to meet, exchange ideas and techniques, form new collaborations, and through the resulting synergetic effects, achieve a step change in progress towards the challenging problems the world faces in terms of managing the textual and visual data deluge.

The aim of iV&L Net is to coordinate ongoing language-vision research within Europe, in order to create a strong and lasting interdisciplinary research community which will address pressing language-vision challenges such as those outlined above. The Action is of crucial importance, timely, and addresses issues not covered, to the best of the knowledge of the initiators of this Action, by any other COST Action, or other current and recent European projects.

iV&L Net will maintain close links with one FP7 project on joint cognitive modelling of action and language, and one video retrieval shared-task competition initiative, which have aims related to one of the Actions core themes.

Bringing the networking and collaboration opportunities, concerted research focus, and community development facilitated by the COST Framework to bear, iV&L Net will achieve substantial scientific and technological impact over the four years of iV&L Net as a COST Action and beyond,

in terms of both upstream advances in fundamental methodologies and downstream progress in technological solutions in the fields addressed.

B.2 Current state of knowledge

Since the origins of Computer Vision (CV), the major, general goal has been image and video understanding. In general, it aims to translate video sequences into high-level semantic concepts (Nagel 1988). Unfortunately, the semantic interpretation of visual evidence for video understanding is not trivial: there is an inherent ambiguity between a sequence of images and its possible interpretations, the semantic gap (Smeulders et al, 2000). In order to bridge this gap, it has been proved useful to rely on the semantic models used in NLP, which aim to detail the essential lower-level attributes of the high-level terms of interest, like ontologies, Case Grammar, Lexical Conceptual Structures, Thematic Proto-Roles, WordNet, Aspectual Classes, or Verb Classes (Ma and Kevitt, 2004). So the use of event modelling tools inspired in NLP became central to understand video content in many applications like smart surveillance, advanced user interfacing, and semantic video indexing.

In addition to this generic image interpretation goal, several CV subtasks have also found inspiration in the NLP domain, e.g. image classification, object segmentation, pose estimation, and motion recognition, among others. Consequently, NLP methods based on bag-of-words, parsing techniques, stochastic grammars or Markov models, to cite but a few, have also been applied successfully in the CV community. As a result, CV typically addresses the above subtasks by constructing symbolic representations (including both appearance and geometry-based ones) through processes variously involving segmentation (possibly expectation-driven), feature extraction and aggregation, and perceptual inference. These representations can then be generalized to yield spatio-temporal categorical and behavioural characterizations by grounding the semantics in new visual experiences through a variety of machine learning techniques. Arguably, language understanding proceeds in the same manner so there are interesting, and exploitable, parallels here.

For image classification, the multimedia and CV communities started using textual tags of image data to supplement image analysis and to help bridge the semantic gap that visual features alone cannot easily fill. Previous research has looked at retrieving visual data based on textual information (e.g. Jeon et al., 2003; Berg et al., 2004; Jain and Learned-Miller, 2007). Research has mainly focused on linking names (recognized with named entity recognizers) from texts with faces detected in image and video data (Everingham et al., 2006; Pham et al., 2010), where most of the work is done in the CV field. In order to understand language, an intelligent system must be able to

connect words, phrases, and sentences to its perception of objects and events in the world. The perceptual context provides the necessary supervisory information, and learning the connection between language and perception grounds the system's semantic representations in its perception of the world.

Taking inspiration from methods originally used in text processing, algorithms for image labelling, search and retrieval have been built upon the connection between text and visual features. Barnard et al. (2003) present one of the first attempts to model multimodal sets of images with associated text, learning the joint distribution of image regions and concrete concepts. Their model has been recently extended to attributes (such as yellow or striped) (Berg et al., 2010, Fahradi et al., 2009, Lampert et al., 2009, Wang et al., 2009), enabling transfer learning to recognize attributes without hand-labelled training data and even new object recognition. Rohrbach et al. (2010) enhance transfer learning for attribute-based classification by using semantic relatedness values that they extract from textual knowledge bases. Both Farhadi et al. (2010) and Kulkarni et al. (2011) aim to associate natural language descriptions to images. They first use visual features to predict the content of an image in terms of objects and attributes. Then they use a natural language generation system to create image captions. Jamieson (2010) propose an algorithm to simultaneously learn the names and the appearances of the objects represented in an unstructured collection of images containing a variety of objects within cluttered scenes.

For image segmentation, recent works assume that image regions have meaning and can be represented by predefined semantic objects or classes. Under this assumption, pixels are grouped into bigger regions based on features in common, and the most popular techniques are the Markov/Conditional Random Fields (Gould et al, 2008) and Random Forests (Shotton et al, 2008). Well-known NLP techniques related to grammars and semantic structures have been also applied in the CV domain for object detection and pose estimation in images, since the seminal paper of Fischler and Elschlager (1973) until the most popular object detector up to date from Felzenszwalb et al (2010). Most state-of-the-art methods are based on these two papers, in which an object is considered as a set of independently learned parts that have spring-like geometrical constraints. Objects are then represented by structures like parsing trees, strings or graphs. For motion recognition, in CV the stochastic evolution of body parts movement over time was initially modelled using Hidden Markov Models, see Poppe (2010) for a complete review. More recently, instead of modelling the motion, action recognition has also been tackled using a bag-of-words model, where the key-poses are obtained using some clustering on limb dynamics (Weinland et al, 2011).

Another interesting line of research exploits the connection between text and images with the goal

to enhance human-robot interaction. For example, Chen and Mooney (2011) present an automatic system that understands natural-language navigation instructions by transforming them into an executable navigation plan. For the task, a semantic parser for interpreting the navigation instructions is learnt by observing how human followers act. Matuszek et al. (2012) present instead a model for grounded attribute learning. Using joint textual and visual information, they build a system capable of producing a set of (visual) attribute models that help in an object selection task. On the NLP side, recently there has been much interest in enriching corpus-based models of word meaning with features extracted from pictures with automated image analysis techniques, to develop a more nuanced and grounded view of word meaning (Feng and Lapata, 2010, Leong and Mihalcea, 2011, Bruni et al. 2012).

In summary, it is clear that the CV community has been inspired by NLP and this benefit will continue, since the impact of NLP techniques in the performance of image processing and computer vision methods has been demonstrated as quite important. The unique innovation of iV&L Net is our focus on the mutual grounding of visual and linguistic meaning, leveraging recent advances that have shown the deep bi-directional interdependence of perception and action in the context of human cognition (Fadiga et al., 2009; Fazio et al., 2009; Iacoboni, 2009; Rizzolatti and Craighero, 2004).

The disciplines can each benefit from the other, in a true symbiosis, and the Action will enable this mutual leveraging. As of now, the state-of-the-art in both fields is that there are only isolated pockets of such mutual leveraging and those that do exist are showing fruitful progress, as described above. But combining expertise and techniques from CV and NLP is receiving considerable interest, as evidenced by several recent workshops held worldwide, such as the CIAM (Cross-media Information Access and Mining) workshop, the workshop on Integrating Language and Vision at NIPS 20011, the Workshop on Language for Vision at CVPR 2013 and the Workshop on Vision and Language (WVL) 2013 held at NAACL HLT 2013.

Clearly, vision is one of the most advanced perceptual faculties and language – specifically speech acts – is perhaps the ultimate form of adaptive action in which humans engage. iV&L Net brings these heretofore separate strands of natural and artificial cognition together in a single operational framework to identify the fundamental research that needs to be carried out in order to achieve real progress and deliver systems that are capable of effectively exploiting the deep semantic interdependence of visual perception and linguistic expression.

B.3 Reasons for the Action

The joint processing of visual and textual data forms an emerging scientific field; a COST Action at this point in time will help ensure that scientific findings can percolate to all potentially interested European research groups. The COST Action brings together European experts in CV and NLP who already have expertise in the subject area of the Action. The Action will reach to the 9 COST countries that have already signed up for iV&L Net, including many of the internationally leading researchers in CV, NLP, Search and Information Retrieval. With the Action an effective framework will be provided to highlight the key research issues, agree effective strategies for addressing them, and share the insights that result from the subsequent coordinated research efforts of the members of the network.

There is a wide range of important real-world application areas that involve both vision and language, including but not limited to: image and video search and retrieval, accessibility tools, human-robot interaction, human-computer interaction in virtual worlds, and computer graphics generation. In Part A three example applications were mentioned which illustrate the real societal need for progress in these areas: medical image processing to cope with the massive amounts of medical images produced daily; CCTV video analysis to enable police to routinely use CCTV footage in combating crime; and automatic generation of descriptions of visual content, to help improve accessibility of websites to visually impaired users.

The above application areas deal with truly ‘big data’. Making sense of big data has huge economic importance. The McKinsey 2011 report on big data describes the mining of such data as the next frontier for innovation, competition and productivity in a variety of sectors including retail, manufacturing and healthcare.

The Action does not merely aim to apply existing technology to the above application areas, but to break new scientific/technological ground through exchange of knowledge and fostering of new collaborations, leading to new research directions.

The Action will result in the following outcomes for its membership, the wider academic community, and industry: (i) channels of communication will be opened up between researchers primarily working on vision problems and researchers primarily working on language problems, thereby facilitating exchange of knowledge and resources; (ii) academic and industrial partners will be brought together for inter-disciplinary collaborations, resulting in new projects and initiatives; (iii) the critical mass of research aimed at vision-language problems, and therefore numbers of projects, publications, and researchers working in the area, will be increased; and (iv) scientific/technological breakthroughs in vision-language challenges including search, retrieval, annotation, description and accessibility of visual data will result.

B.4 Complementarity with other research programmes

This Action is unique in bringing together Natural Language Processing (NLP) and Computer Vision (CV) research communities at a European and international level. There has been only one effort at national level for bringing together these two fields. iV&L Net builds on this effort and the valuable lessons and experience acquired through this regional initiative; these suggest a pan-European Action, with international character and resources that will allow for true collaboration between participants. The regional experience provides the important background for an informed, well-organised kick-off for iV&L Net with high potential for cross-fertilization of its different computation-oriented research communities. As far as can be ascertained, no similar effort is supported by COST or any other funding mechanism worldwide.. As discussed in Section E.3, there are some COST Actions which are indirectly related to iV&L Net themes, and with which collaboration will be fostered (e.g. MUMIA, and ENERGIC).

Furthermore, there are a number of research and development projects, currently funded by the European Commission, whose objectives are related in one way or another to vision and language integration issues, and in that sense they are also indirectly linked to activities in iV&L Net (for example, FP7 POETICON++ ICT-2011.2.1 Cognitive Systems and Robotics , FP7 NOVICOM FP7-PEOPLE-IEF-2008, FP7 SAVAS ICT-2011.4.1 SME initiative on Digital Content and Languages, FP7 APIDIS ICT-2007.4.2 Intelligent content and semantics, Cognimund ERC-SG-PE6 ERC Starting Grant, FP7 MUSE Machine Understanding for interactive StorytElling ICT-2011.9.1 Challenging current thinking). Similar projects are funded in other parts of the world too (e.g. USA). iV&L Net will liaise and interact with these projects as appropriate.

There is a long history of individual projects with indirect interests in vision and language integration going back to the very early days of Artificial Intelligence. However, none of these projects addressed this core issue as such. They all focused on a variety of applications that require vision-language integration mechanisms, without actually addressing the issue of vision-language integration directly. This is indicative of the need for focussing research activities on the automatic integration of language and vision, which is at the core of this Action. Thus, iV&L is not only complementary to such previous research efforts, but also leads in the forefront of core research issues in this area, emphasising the need for research coordination, aiming at producing innovative ideas and providing a flexible environment for interdisciplinary research.

C. OBJECTIVES AND BENEFITS

C.1 Aim

The main objective of the Action is the creation of an interdisciplinary European vision-language research network targeting scientific advances and societal benefits in the four focus themes of (i) Integrated Modelling of Vision and Language for Computer Vision and Natural Language Processing Tasks; (ii) Applications of Integrated Models; (iii) Automatic Generation of Image and Video Descriptions; and (iv) Semantic Image and Video Search.

C.2 Objectives

The overall expected impact of iV&L Net is to bring about a step change in progress towards effective solutions for computational challenges involving both language and visual content. In particular, iV&L Net will focus on the integration of language and vision, and the benefits this brings to such solutions. Our specific objectives are:

1. To advance theory, methodology and real-world technology across the language-vision spectrum, but particularly in the four iV&L Net focus research themes: (i) Integrated Modelling of Vision and Language for CV and NLP Tasks; (ii) Applications of Integrated Models; (iii) Automatic Generation of Image and Video Descriptions; and (iv) Semantic Image and Video Search.

2. To facilitate collaboration, networking and interdisciplinary community building;

3. To liaise extensively with industry and end-users;

4. To coordinate the development of benchmark data resources for tasks relating to the focus themes above;

5. To specify grand challenges and organise corresponding shared-task competitions relating to the focus themes.

C.3 How networking within the Action will yield the objectives?

The Action will enhance the overall scientific synergy between the disciplines of Computer Vision and Natural Language Processing. The different types of knowledge and expertise brought together by the Action will cross-fertilize algorithmic thinking, and will bring about novel and fresh ideas on language/vision challenges.
The following aspects of networking within the Action will directly contribute to bringing about the objectives in C.1 and C.2:

1. Exchange of resources including semantic annotation schemes and guidelines, benchmarking corpora, machine learning and alignment tools.

2. Electronic tools to facilitate interactions, collaborations, knowledge building and dissemination: a website that (i) provides information about members, their activities and contact details, the ongoing research coordinated by the Action, conferences, dissemination activities, and training opportunities; (ii) facilitates participation in an iV&L Net blog and in forum discussions, and (iii) makes important publications available for download.

3. Organization of training schools including summer schools and conferences focused on the generation of novel ideas and on introducing researchers to the new joint CV-NLP discipline.

4. Exchange visits of Early Stage Researchers (ESRs) especially focusing on NLP researchers visiting CV labs and vice versa, and other activities that encourage young researchers to establish links with industry and more senior academics.

5. Scientific dissemination through the above mentioned conferences, scientific and industrial gatherings which will have substantial impact in the participating countries and beyond.

6. Scientific publications in books, journals and conference proceedings; reports from WG meetings and training schools; training materials.

7. Joint applications for European and national funding for research projects within the fields covered by the Action, in order to encourage novel outcomes and establish critical mass.

C.4 Potential impact of the Action

For the scientific community the prime tangible benefits of the Action will be (i) advanced techniques and tools capable of more accurately processing, indexing, searching and mining language and visual content, (ii) resources and annotation guidelines relating to such content, and (iii) cross-fertilisation of the machine learning, pattern recognition and semantic processing algorithms currently in use in the CV and NLP communities. Moreover, the Action will provide access to other representatives from academia, industry and public institutions facilitating exchange of knowledge and expertise and fostering new collaborations.

The Action will have substantial impact for society. The technologies that iV&L Net will be working towards have the potential to benefit a wide range of different users. People with impairments in sight, hearing and cognitive ability will benefit from assistive technology that will improve accessibility in an increasingly multimodal world. Improvements in image search and retrieval will enhance online search experience, as well as help institutions such as hospitals and police forces to cope with ever growing quantities of images and videos. Better understanding of natural language by integrating background common knowledge obtained from the visual medium has enormous application in language understanding (mapping language to knowledge to be used e.g. in decision making) and access to textual content (e.g., bringing text to life in a virtual world for persons with reading impairments).

With regard to economic impact, the software and services that next-generation language-vision technology will feed into have huge commercial potential. Search engine providers are competing to improve image search, as the keyword matching techniques currently employed have clear limitations. The Action will lead to advanced CV, NLP and multimedia applications. There are numerous European companies that specialise in assistive technology which would be interested in taking new assistive tools to market. An Industry Advisory Group and End-user Advisory Groups will be closely involved in all stages of iV&L Net, including research proposal preparation.

C.5 Target groups/end users

The target groups and end users of the Action include:

1. Companies and institutions who have the management of large multimedia archives as part of their remit, including e.g. the multimedia industry which has a number of large companies in Europe, and companies specialising in search and retrieval.

2. Companies and professionals who develop tools for CV, NLP and multimedia processing.

3. Academic researchers and educators working on textual and visual content processing and management. The Action's proposing team includes ten leading academic researchers with strong track records in their specific fields.

4. The European Commission who will be able to use results from the Action to inform policy making and call content selection.

5. Young researchers who will be trained in the interdisciplinary field addressed by the Action through scientific exchange visits and training programs. Support and involvement of ESRs will form an integral part of this Action.

6. End users who will benefit from diverse advanced tools resulting from the Action, facilitating e.g. advanced semantic search, improved access to visual and textual data, and possible translation from one modality to the other.

D. SCIENTIFIC PROGRAMME

D.1 Scientific focus

The research work to be carried out by iV&L Net participants at national level, and coordinated by iV&L Net, is structured around four focus themes and will target the following chief aims:

1. Integrated Modelling of Vision and Language for CV and NLP Tasks: to coordinate and build on ongoing research on integrated L/V models, including cognitive modelling

approaches and probabilistic synchronous grammar models, and to share the associated expertise and tools among the network;

2. Applications of Integrated Models: to apply integrated V/L Models to a variety of applications involving both language and vision, and to test their performance on benchmark data in competitive situations;

3. Automatic Generation of Image and Video Descriptions: to integrate image/video analysis techniques with language generation methods in order to achieve a breakthrough in automatic annotation and description of visual content, both for the purpose of search/retrieval and for accessibility of visual content by blind and partially sighted people; and

4. Semantic Image and Video Search: to systematically bring language models, word nets, and other linguistic technologies to bear, in conjunction with the most advanced image/video analysis, on the task of identifying semantic content in images/video.

In order to facilitate collaboration, networking and interdisciplinary community building iV&L Net will introduce a variety of coordination and collaboration mechanisms (see Section C.3). Section D.2 below describes the research work to be carried out at Working Group (WG) level and coordinated by the Management Committee (MC). Section E.1 provides a list of milestones, and Section F a list of networking events and activities, and a timetable to show when over the four years of the Action they occur.

This Action's timeliness is clear from several recent national initiatives and workshops aimed at interdisciplinary vision and language research; the Action is highly innovative in its international dimension and specific scientific goals including the focus on the integration of vision and language, and in being supported by multi-national collaboration, coordination and networking mechanisms. The Action’s ambitious goals will be facilitated by a range of technical means as listed in Section C.3, and by the MC and WG roles and tasks that have been defined and tailored to the requirements of this Action (as given in Section E.1).

D.2 Scientific work plan methods and means

The principal tasks for iV&L Net are to identify the scientific developments needed to bring about vision and language integration (road-mapping), to stimulate development of new integrated approaches (scientific/technological foundations), to demonstrate the successful application of the integrated approach (applications), and to act as a catalyst in all of the above. In short, coordination and support is required to enable the advancement of the discipline and to lay the foundations for making it happen. This is the goal of iV&L Net. The Action's focus is emphatically not just on the network as an organization, but also on the network dynamics: the collaboration, the partnering, and interaction facilitated by the research network, with a commitment to produce tangible results. Coordinated research under iV&L Net is structured into five Working Groups (WGs) each with content and tangible outputs as described below. All WGs are supported by the following types of Action activities:

1. *Advancing Science* activities will address research planning and actions focused on consolidating the existing body of knowledge, as well as identifying what needs to be done in order to advance the state of the art. Considerable effort will be devoted to consolidating the fragmentary body of knowledge underlying the network at its inception, to serve as a basis for creating an ambitious and comprehensive research agenda. The key focus here is the cross-fertilization of ideas from many areas through the organization of thematic, consolidation and research roadmap workshops.

2. *Education & Training* activities are intended to address the difficulties posed by the multi-disciplinary nature of the area. The goal is to provide an effective mechanism to bridge gaps between sub-disciplines and help researchers in one area come up to speed in other areas. It will be targeted both at researchers and graduate students.

3. *Shared Resources for the Community* activities focus on providing a web-based repository of material (accessible via the iV&L Net website) that will assist the iV&L Net community, in research and in education; the repository will also increase the visibility of the relevance and importance of integrated vision and language technology to the wider research community beyond iV&L Net.

WG1 Integrated Modelling of Vision and Language for CV and NLP Tasks

WG1 focuses on core theoretical approaches to joint vision and language processing by developing

vision and language models that are integrated to various degrees. Important tasks will be to consolidate existing knowledge, define annotation standards, and draft a roadmap. The complexity of the area (due to its currently fragmented make-up and its emerging nature), calls for an innovative approach tightly focused on scientific/technological progress.
WG1 Tangible Outputs:

- Semantic annotation guidelines and standards for vision and text.

- A repository of open source software that covers valuable processing software for language and visual content including a directory of sources of materials or components, with specifications.

- Survey and position papers.

- Organization of training and summer schools; their course wares and bibliographies, including short courses on CV and NLP focused on accelerating cross-topic learning for iV&L members, and courses for the wider community.

- Curriculum for a graduate course on the cross-modal processing of vision and text.

- A cross-topic research roadmap that will identify where research effort is most needed and the research challenges that need to be addressed.

WG2 Applications of Integrated Models

WG2 focuses on specific applications. This WG activity embraces both intra-Action collaboration and the involvement of new participants; collaboration between academic and industrial partners, both on academic projects and on real-world product development. The activity will stimulate ideas for novel cross-modal applications. It will include initiatives for bi-lateral exchanges, particularly for individuals who are not yet directly involved with funded projects located in the iV&L Net area, in this way addressing the wider community. Among the methods applied are intra-Action workshops and extra-network workshops that establish contacts with European and nationally funded projects.

WG2 Tangible Outputs:

- Requirements analyses for breakthrough progress in a variety of application contexts.
- Repository of application-oriented demonstration scenarios to drive R&D.
- Demonstrators for given applications accessible via the iV&L Net website.
- Articles for general readership.
- Project proposals at European and national level.

WG3 Automatic Generation of Image and Video Descriptions

WG3 focuses on the methods for annotating, labelling and describing visual data, including integration of language technologies in annotating visual data using suitable weakly supervised machine learning models, inference models that take into account language and visual constraints, latent class models for coping with variant low level features, alignment models, models that detect complementarity of the vision and text content, and text generation methods.

WG3 Tangible Outputs:

- Construction of benchmarking datasets for an application that targets visually impaired people.
- In-depth analysis and review of methods for annotating visual data based on weakly supervised learning from text.
- Development of algorithms and strategies potentially resulting in a standard methodology for the analysis of visual data.
- Organization of an international shared-task competition workshop.

- Construction of a repository of demonstrators.

- Exchange of staff and students, Short Term Scientific Missions (STSMs).

WG4 Semantic Image and Video Search

This WG applies insights from integrated vision and language processing to the important application of video and image search, and focuses on suitable retrieval models that reason with and fuse the results of uncertain recognitions.

WG4 Tangible Outputs:

- In depth analysis and review of methods for semantic search of visual data.

- Development of retrieval model best practices and guidelines.

- Participation of iV&L members in existing image and video search competitions (e.g. TRECVid).

- Exchange of staff and students, Short Term Scientific Missions (STSMs).

- Organization of a course at an information retrieval summer school, courseware and bibliography.

WG5 Industry and End-User Liaison

This WG will be responsible for developing links with industry and end users. The Action will invite larger and smaller companies who are stakeholders in the areas addressed by iV&L Net to join an Industrial Advisory Board which has the dual function of (i) advising the Action's MC and informing its activities, and (ii) fostering collaboration between industrial and non-industrial participants, including industry placements for ESRs.

WG5 will also coordinate the recruitment of End-user Advisory Groups, including e.g. one of blind and partially sighted internet users, and one of video analysis stakeholders such as police forces and CCTV companies. Furthermore, WG5 will coordinate user requirement surveys and other methods for obtaining end-user input.

WG5 Tangible outcomes:

- Establishment of an Industrial Advisory Board.

- Organisation of Industrial Placements.

- Establishment of End-user Advisory Groups.

- Reports on requirements surveys.

E. ORGANISATION

E.1 Coordination and organisation

The Action will be coordinated by the iV&L Net Management Committee (MC) formed and acting according to COST Rules and Procedures. MC Chair, Vice-Chair and Secretary will be elected at the Action's Kick-off Meeting. Chair and Vice-Chair will preside over the MC and oversee the work of the five WGs, each of which is managed by a WG Coordinator and Vice Coordinator. The MC will hold two meetings per year, in conjunction with the meetings of the WGs. There will be three further MC roles, Scientific Coordinator, STSM Coordinator and Dissemination Coordinator, also to be elected at the Action's Kick-off Meeting.

The Scientific Coordinator's tasks will include periodically reporting to the MC on the scientific progress of the Action; overseeing the implementation of a shared repository of materials/work; liaising with the WG Coordinators about work to be done; providing advice on research topics based on the MoU; in collaboration with the WG Coordinators, developing scientific/technical programmes for Action Workshops; together with the Dissemination Coordinator, organising the publication of a series of books/proceedings resulting from the Workshops/Conferences; and generally overseeing the scientific activities of the Action, identifying any need for improvement/discussion at MC Meetings.

The STSM Coordinator will be in charge, together with the WG Coordinators, of coordinating STSMs; to oversee the process whereby applicants are selected for STSMs, including calls, selection criteria, reviewing, and ensuring required ESR representation (at least 50%).

The Dissemination Coordinator's tasks will include forming a Dissemination Committee; drafting a Dissemination Plan in the first quarter of the Action; overseeing the design and production of dissemination materials; coordinating the promotional activities for Workshops/Conferences to

ensure broad participation; liaising with key industrial and academic partners; periodically publishing an Action Newsletter; maintaining a regularly updated list of Action Participants, stakeholders, end-users and other target audiences and keeping them informed about the Action's activities; generally overseeing the dissemination activities of the Action and identifying any needs for improvement/discussion at MC Meetings.

Coordination of research will be focused around the four focus research themes outlined in Part C, each overseen by a dedicated Working Group (WG). A fifth WG will be responsible for developing links with industry and end users. The 5 WGs will be responsible for coordinating national research under each of the 5 headings. WG 5 will be supported by the Industrial Advisory Board of industry representatives.

iV&L Net will benefit from an interactive website offering access to a database of members and their expertise, searchable to members; online 'matchmaking' services for members looking for collaborators with specific expertise, including a dedicated academia-industry matchmaking service; a repository of data resources; a searchable repository of software tools and resources; a mailing list where members can post questions and announcements; and a periodic newsletter.

iV&L Net will set up an Industrial Advisory Board, organise Annual Conferences, Technical Meetings, Industrial Secondments and Partner Visits (50% reserved for Early Stage Researchers), and have a flexible membership structure. iV&L Net will produce periodic reports on the state of the field and progress achieved, coordinate data creation and shared-task competitions in the core research themes, and conduct a road-mapping exercise in the vision and language interdisciplinary field.

Europe has many of the world's leading researchers in NLP and CV. iV&L Net will tap into this body of expertise to create new strategic partnerships aimed at narrowing the language-vision gap by developing the theory required for solutions to the difficult challenges posed by an increasingly multimodal world. A successful network will place Europe at the forefront of developing language-vision solutions with clear commercial potential. A COST Action with its support for coordination of research, networking among stakeholders, exchange of expertise, and potential for forming new research partnerships, is exactly the right framework for the aims and activities outlined in this document, and this particular form of support is not available under other funding schemes.

Milestones:

M1 [end of Month 3] The Action is Up and Running: All Planning Documents (Dissemination Plan, WG Work Plans, ESR Action Plan, Gender Balance Direct and Indirect Action Plans, etc.) are available, initial WG recruitment drive is complete, Action Website is up and running, First MC Meeting has taken place.

M2 [end of Year 1] MC and WG Targets and Deliverables for Year 1 Have Been Achieved: Reports/publications from the workshops and conferences held; establishment of methodologies for jointly processing vision and text reflected in reports and training courseware; benchmarking data collections, shared-task definitions, and required annotations for one selected application are available.
M3 [end of Year 2] MC and WG Targets and Deliverables for Year 2 Have Been Achieved: Reports/publications from the workshops and conferences held; establishment of methodologies for jointly processing vision and text reflected in reports and training courseware; benchmarking data collections, shared-task definitions, and required annotations for two selected applications are available; first set of tools and demonstrators covering selected applications are available.
M4 [end of Year 3] MC and WG Targets and Deliverables for Year 3 Have Been Achieved: Reports/publications from the workshops and conferences held; establishment of methodologies for jointly processing vision and text reflected in reports and training courseware; benchmarking data collections, shared-task definitions, and required annotations for three selected applications are available; expanded set of tools and demonstrators covering selected applications are available.
M5 [end of Year 4] MC and WG Targets and Deliverables for Year 4 Have Been Achieved: Final conference has been held; reports/publications from the workshops and conferences held; establishment of methodologies for jointly processing vision and text reflected in reports and training courseware; benchmarking data collections, shared-task definitions, and required annotations for three selected applications are available; final set of tools and demonstrators covering selected applications are available.

E.2 Working Groups

The Action will be structured around five Working Groups (WGs). Work in each WG will be coordinated by its WG Coordinator who will be appointed during the Action's Kick-off Meeting. The WG titles are as follows; their content is described in more detail in Part D. Membership in different WGs will be open to all participants and not mutually exclusive.

1. Integrated Modelling of Vision and Language for CV and NLP Tasks;

2. Applications of Integrated Models;

3. Automatic Generation of Image and Video Descriptions;

4. Semantic Image and Video Search;

5. Industry and End-user Liaison.

Each WG will be led by its WG Coordinator and Deputy Coordinator. They will (i) organize and chair WG Meetings, prepare meeting agendas and meeting minutes; (ii) coordinate and review scientific/technical work; (iii) ensure continuous/efficient communication within and across WGs; (iv) periodically prepare reports to the MC; and (v) communicate regularly with the Scientific Coordinator and Dissemination Coordinator on progress.
In the first quarter of the Action, each WG will recruit participants from the wider CV and NLP communities including industry, public institutions, and the Actions and projects mentioned in Sections B.4 and E.3. Also in the first quarter, each WG will draft a WG Work Plan, identifying a WG specific work programme, deliverables and performance indicators.

E.3 Liaison and interaction with other research programmes

iV&L Net will liaise and interact with European resource building programmes such as META-NET (FP7 249119, 271022, 270893 and 270899) and CLARIN (FP7 212230) for natural language resources, in particular in relation to the benchmark challenges and associated data that will be created in iV&L Net.
While there are no COST Actions located in the same fields as iV&L Net, there are some that address issues tangentially related where interaction will produce synergetic benefits. E.g. iV&L Net will consider interacting with MUMIA (Multilingual and Multifaceted Information Access) which covers the areas of Machine Translation (MT), Information Retrieval (IR) and Multifaceted Interactive Information Access (MIIA). Participating groups of MUMIA have developed technology for term alignment in comparable datasets. Reaching out to MUMIA could cross-fertilize content alignment in multilingual and multimodal data.
ENERGIC (European Network Exploring Research into Geospatial Information Crowd-sourcing) explores new and unprecedented sources of geographic information in the form of user-generated Web content. The iV&L Net technologies will offer ways to automatically annotate images and videos captured at multiple spatial and temporal scales, so ENERGIC could be a valuable source of

end users for iV&L Net.
iV&L Net will liaise with the above and other initiatives and research and development projects (such as POETICON++, NOVICOM, SAVAS, APIDIS, Cognimund, see Section B.4 for more details) for example by (i) regularly informing the respective coordinators on the Action's scientific programme and progress; (ii) co-organising, whenever appropriate, common events such as workshops within international conferences of the domain; and (iii) promoting the Action's Short Term Scientific Missions and Training Schools among Early Stage Researchers included in these projects and actions.

E.4 Gender balance and involvement of early-stage researchers

This COST Action will respect an appropriate gender balance in all its activities and the Management Committee will place this as a standard item on all its MC agendas. The Action will also be committed to considerably involve early-stage researchers. This item will also be placed as a standard item on all MC agendas.
Computer Science as a whole currently has low, and declining, levels of participation from women. While in the 1980s just over a third of computing science degrees were awarded to women, this number has now shrunk to below 12% (2010/12 Computing Research Association Report, US). Interestingly, Vogel and Jurafsky (2012) found that the percentage of women authors in the subfield of Natural Language Processing has been continuously increasing since the 1980s, approximately doubling during this time. Compared to the Computing Research baseline of 12%, participation of women in the proposing team of this COST Action is respectable at 21.43%.
In the course of the first year of the Action the aim will be to increase this proportion to one third (to match the historical topline) on the MC, and to one quarter on average in the Working Groups (WGs). The latter is an ambitious aim, but one which is nevertheless realistic, given the higher than average participation of women in NLP. In addition to the direct action aimed at achieving given levels of participation in the MC and WGs, iV&L Net will also engage in indirect action aimed at promoting best practice in improving gender balance in recruitment of researchers among Action participants (in this, the Code of Conduct for Recruitment of The European Charter and Code for Researchers will be followed). To these ends both a Direct Action Plan and an Indirect Action Plan will be drafted in the first quarter of the COST Action, detailing specific steps and targeted outcomes.
This COST Action is aimed at a radically new research direction---bringing the combination of vision and language to bear on urgent research challenges---and as such will crucially rely on the

involvement and enthusiasm typical of Early Stage Researchers (ESRs). ESRs will be involved at every level of iV&L Net, from the MC and WGs to meetings and training schools, the latter being specifically targeted at ESRs. The Action will aim for a minimum of 20% ESR participation in the MC, and 40% in the Working Groups. Partner Visits will be 50% reserved for ESRs. An ESR Action Plan will be drafted in the first quarter of the COST Action, detailing specific steps and targeted outcomes.

The above percentages (subject to confirmation in the Direct and Indirect Action Plans) constitute the Performance Indicators against which success in achieving balance with regard to gender and ESRs will be measured.

F. TIMETABLE

The duration of the Action is 4 years and the table below provides an overview of the following activities to be undertaken.

MC meetings: 2 MC meetings will be organized annually at least 1 of which will be co-located with other activities of the Action.

WG meetings: Each WG will organize 8 WG meetings, the WG meetings at the end of Year 2 and Year 4 will be organized jointly. The Year 1 and Year 3 WG meetings will be co-located with an international conference in the Action's area.

Training schools: A short training session for the Action participants will be organized at the start of the Action jointly with the kickoff meeting. Two subsequent training schools will be organized, in Year 2 and Year 4, focusing on the interaction between CV and NLP.

Workshops and final conference: Three open access workshops will be organized in conjunction with international conferences in Year 1, Year 2 and Year 3 of the Action. During the final open access conference at the end of the Action, international experts whose research is relevant for the Action will be invited.

Other dissemination activities: Dissemination via the public website will start immediately and communication/project management tools will be installed on the project-internal website.

STSMs: Throughout the Action, STSMs will be set up within and between WGs. At least four STSMs per year will take place to stimulate collaboration between the CV and NLP fields.

	Year 1	Year 2	Year 3	Year 4

	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
MC meetings	x		x		x		x		x		x		x			x
WG meetings	x		x		x		x		x		x		x			x
Training schools							x								x	
Workshops/ conferences			x				x				x					x
Other disseminations	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
STSMs	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

For list of milestones, see Section E.1.

G. ECONOMIC DIMENSION

The following COST countries have actively participated in the preparation of the Action or otherwise indicated their interest: BE, CH, EL, ES, IE, IT, NL, SE, UK. On the basis of national estimates, the economic dimension of the activities to be carried out under the Action has been estimated at 36 Million € for the total duration of the Action. This estimate is valid under the assumption that all the countries mentioned above but no other countries will participate in the Action. Any departure from this will change the total cost accordingly.

H. DISSEMINATION PLAN

H.1 Who?

As is the norm for a COST Action, dissemination activities will target the research, teaching and industry institutions that participate in the Action, targeting in particular Early Stage Researchers (ESRs), and including Master and PhD students, who will be able to exploit all educational and networking opportunities offered by the initiative. While it may seem strange to have such inward-facing dissemination, what is novel about this COST Action is that the Action's community of researchers does not have a history of collaboration, or even of mutual awareness, thus the within-Action dissemination is not downplayed and is regarded as an important goal. In particular, the training events within the Action will go a long way towards harmonising the work within the Action, by raising mutual awareness of all disciplines contributing to the topic.

As well as inward-focusing the Action also targets outward-focusing dissemination to the broader academic community in the fields of natural language processing, computer vision, information retrieval and human-computer interaction (specifically, computer-based assistive technologies) as well as to the broader ICT industry community.
Outside of the obvious academic / researcher groups within and outside the Action, public bodies that are mandated to ensure universal accessibility to multimodal resources are an important focus of the Action outreach and can benefit directly from the Action in how they achieve their own mission.
Public institutions with large multimodal data management needs including national archives as well as the niche multimedia collections in hospitals, police forces, etc. can exploit the developments expected from within the Action and participants in the Action already have ongoing collaborations with such institutions.
The multimedia industry has a number of large companies in Europe who have the management of large multimedia archives as part of their remit, and these will be a focus for targeted dissemination with an emphasis on exploitation.
Finally, our dissemination plan will also have a component aimed at the general public and the media in order to raise awareness of what can be done with content-based access to rich multimedia information.

H.2 What?

The Action website will be the main dissemination tool, offering access to a database of members and their expertise (restricted to members); online 'matchmaking' services for members looking for collaborators with specific expertise, including a dedicated academia-industry matchmaking service; a public repository of data resources; a public searchable repository of software tools and resources; public archives of all the publications and reports produced by the Action.
The website will also contain links to a portal of partners' training materials which will have been delivered at various Action training events for educational and instructional purposes, as well as emanating from within the Action consortia membership, and this will be stored on 3rd party sites.
The Action will activate a mailing list where members can post questions and announcements, also open to other interested researchers and professionals.
The Action will issue a periodic newsletter with a wide distribution to academic and industry partners as well as to European politicians and stakeholders in the European Commission.
The flow of communication and collaboration among Action participants will be enhanced by

technical meetings and partner visits (50% reserved for Early Stage Researchers) which will be the most costly of the dissemination activities.
iV&L Net will produce periodic reports or 'white papers' on the state of the field and progress achieved. These will be collaboratively authored with inputs from across the Action and the co-authoring process itself will help to cement relationships, increase mutual awareness and create a sense of the collaborative nature of the COST Action.
The Action will benefit from a yearly Vision-and-Language strategic workshop (to take place in at least three calendar years during the Action) that will co-locate and rotate across major conferences in NLP (e.g., ACL), Computer Vision (e.g., ICCV), Information Retrieval (e.g., SIGIR) and Human-Computer Interaction (e.g., CHI), to attract different research communities. This will not be a research event with the usual types of scientific presentations but will be reflective and strategic, bringing communities together to reflect on how the disciplines intersect and interact and how such interaction can be improved and what benefits might accrue.
In association with the workshops, the Action will design shared-task competitions or benchmarking challenges involving the integration of vision and language, that will encourage collaboration between language- and vision-oriented research teams.
While the workshop activities will have a focus on dissemination of high-quality research achievements and outputs, the Action will also run a series of training events, somewhat like Summer Schools but not as intensive as a full 5-day school. These are aimed at the intersection between our contributing disciplines, helping researchers to learn more about the other disciplines, not more about their own.
At the end of the Action, a widely advertised dissemination event will be held targeting industry and public bodies, updating the wider audience about the state of the art and possible roadmaps for the interdisciplinary vision and language field, as well as focusing on concrete applications.
Finally, scientific publications in academic journals and conference proceedings are an effective dissemination device targeting the scientific/academic community.

H.3 How?

Dissemination activities are coordinated by a Dissemination Committee (DC) led by a member of the Management Committee (the Dissemination Coordinator) and representatives of all the Working Groups as well as of industry partners. The DC is appointed at the kick-off meeting, and it will be in charge of devising a Dissemination Plan, including developing and finalising the parameters for the Key Performance Indicators (KPIs), as well as overseeing the implementation of

the plan and revising it, if necessary, on a yearly basis. The performance of the Dissemination Plan will be a standing order item on the agendas of all Management Committee meetings and the progress of the dissemination activities against its KPIs will be closely monitored. This approach allows buy-in from researchers across all the Action partners and ensures dissemination is an integral part of the Action. The combination of technical meeting, partner visits and exchanges, white papers, annual workshop with shared task benchmarking competitions to add focus, training events for ESRs, final dissemination event and scientific publications represents a broad portfolio of dissemination methods which are complementary in nature. Many of these dissemination activities will continue beyond the initial 4-year lifetime of the Action.