INFORMATION SOCIETIES TECHNOLOGY
(IST)
PROGRAMME

Contract for:

Shared-cost RTD

Annex 1 - "Description of Work"

Project acronym: VIBES
Project full title: VIdeo Browsing, Exploration and Structuring
Proposal/Contract no.: *IST-1999-26001*
Related to other Contract no.: *n/a*

Date of preparation of Annex 1: August 2000

Operative commencement date of contract: *see Article 2.1 of contract*

Contents

1. Project Summary

Objectives (maximum 1000 characters)

Video provides continuous coverage of scenes over an extended region both in time and in space. That is what makes it more than a plain collection of images. In VIBES, our objective is to make video a first class data type, which can be searched on content, annotated, hyper-linked, and edited much as text can be now. Furthermore, video
has many more modes of information than simple text. For example, it contains scene geometry and extended actions over multiple frames.

Our objectives are also to extract and use these ``modes''. With these aims, VIBES proposes new ways of exploring and using video that have the potential of leading to significant breakthroughs in video consumption and new industrial, commercial, and home entertainment applications. The tools we develop will enable cut detection, indexing, synthesis, and classification of non-static and non-rigid scenes.

Description of work (maximum 2000 characters)

The project contains eight interlinked workpackages investigating
two main themes:

1. Rapid browsing and retrieval:

A video or a DVD will be automatically augmented with hyperlinks connecting shots containing a particular actor, type of action, or scene. E.g. all scenes inside the casino in `Casablanca'. Such facilities will change the way in which video is addressed, significantly reducing the tedium and inefficiency of current serial video browsing.

2. 3D scene synthesis and human animation models:

3D scene geometry for virtual reality environments will be automatically generated for particular shots. E.g., the yellow brick road in the `Wizard of Oz' could be reconstructed, and a viewer could then walk down it using VR together with virtual actors. 3D dynamical models of actors from classic movies will be learned and used to generate new scenes
involving the actors -- synthetic thespians -- or to replace one actor by another. For example, replace the `hero' in Home Alone or Toy Story by a texture mapped dynamical model of your son or daughter.

Milestones and expected results (maximum 500 characters)

YEAR 1: Simple video unit segmentation. Feature based matching between shots. Simple object recognition. 3D scene models of shots. Simple within-shot human tracking.
YEAR 2: General video unit segmentation. Multi-shot matching. 3D scene models from multiple shots. Advanced human modelling and tracking. Initial hyper-linking demonstrator.
YEAR 3: Classification of action and certain object and scene types. Sequence-to-sequence alignment. Merging human models from multiple shots. Final web-based hyper-linking and video synthesis demonstrators.

WP1		WP2	WP3	WP4	WP5	WP2	WP3	WP2	WP3	WP10	TOTAL
KTH – admin								6	4	6	16
KTH	6		24	1		12		1	1	1	46
Weizmann	34	24	29	4				1	1	1	94
INRIA	6	14	15	4	6	18	5	1	1	1	71
Leuven		31	4	4	12		4	1	1	1	58
Oxford		24	4	4	12		3	1	1	1	50
EPFL						36	7	1	1	1	46
TOTAL	46	93	76	17	30	66	19	12	10	12	381

2. Project Objectives

Video is a rich source of information useful in many aspects of our lives, *e.g.*, for entertainment, surveillance, and professional uses. While standard sequential video is well adapted to viewing in "movie mode", it is very inefficiently organized for applications requiring more active manipulation of video data. The information of interest is buried *implicitly* and *redundantly* inside vast amounts of raw video stream data (approximately 25 Megabytes per second), and is distributed across many frames. So far, these limitations have prevented video from becoming a *manipulable data type* — unlike audio or textual data, which are actively used, manipulated, and browsed over the Internet today. In particular, standard video streams fail to support rapid access to, and object-level manipulation of, objects, people and information of interest in the video, which is needed for many of the emerging applications of video.

The strategic objective of VIBES is **to make video a first class data type**, which can be searched on content, annotated, hyper-linked, and manipulated, much as text can be now. In fact, video contains multiple overlapping "modes" of information, *e.g.* scene geometry, object appearance, extended actions over multiple frames. In VIBES, we will extract and use several of these modes. In this way, VIBES will provide new methods of exploring and using video, which have the potential of leading to significant breakthroughs in video consumption and new industrial, commercial, and home entertainment applications. The tools we develop will enable cut detection, indexing, classification, manipulation and synthesis of non-static and non-rigid scenes.

To achieve this, VIBES has two main technical objectives: to enable **video searching and hyperlinking** based on video segmentation, matching and classification; and to enable **video resynthesis** based 3D scene synthesis and human dynamical modeling.

Video searching and hyperlinking: A video or a DVD will be automatically augmented with hyperlinks connecting shots containing a particular actor, type of action, or scene. *E.g.* all scenes inside the casino in 'Casablanca'. To achieve this, the video will be parsed into primitive entities (spatio-temporal video units) describing scenes, video objects, and actions and motions. These entities will then be matched throughout the video and, to a limited extent, classified. The resulting information will be condensed into a form suitable for element searches and video hyperlinking. The hyperlinked video will provide indexing capabilities analogous to MPEG-7, supporting operations such as element search and video editing and manipulation.

Video resynthesis: 3D scene geometry for virtual reality environments will be automatically generated for particular shots. *E.g.*, the yellow brick road in the 'Wizard of Oz' could be reconstructed, and a viewer could then walk down it using VR together with virtual actors. 3D dynamical models of actors from classic movies will be learned and used to generate new scenes involving the actors ("synthetic thespians"), or to replace one actor by another. The output will be suitable for use in video synthesis modules similar to VRML-2 or MPEG-4.

These objectives involve the following steps over the three years of the project (these are discussed in more detail in §9):

Objectives	**Measure of success**
Year 1	
Simple video unit segmentation.	Independent motion segmentation for a panning camera.
Pairwise scene matching between shots.	Correctly matched shots for test video.
Simple 3D scene models of shots.	3D VRML model from unoccluded shot of a Hollywood film.
Simple within-shot human tracking.	Dynamical model from unoccluded human walking in a Hollywood film shot.
Year 2	
General video unit segmentation.	Test video segmented according to object motion.
Multi-shot matching.	Scene and object matched shots throughout test videos.
3D scene models from multiple shots.	3D VRML model built from several unoccluded shots of a Hollywood film.
Advanced human modelling and tracking.	Human tracking in clutter. Detailed 3D model of human from uncluttered videos.
Year 3	
Classification of action and certain object and scene types.	Video human motion sequences classified into walking/running. Qualitative scene descriptions such as man-made/natural computed automatically for shots of test videos.
Merging human models from multiple shots.	3D human model built from several shots of a Hollywood film.
Web-based hyper-linking demonstrator.	Enables interactive searching and jumping on hyperlinks for test videos.
Web-based video synthesis demonstrator.	Actor and/or scene replacement.

3. Participant List

Role	Partic. No.	Participant name	short name	Country	Date enter project	Date leave project
CO	1	Kungliga Tekniska Hogskolan	KTH	Sweden	start	end
CR	2	Weizmann Institute of Science	WIS	Israel	start	end
CR	3	INRIA Rhône-Alpes	INRIA	France	start	end
CR	4	K.U. Leuven	KUL	Belgium	start	end
CR	5	University of Oxford	OXF	U.K.	start	end
CR	6	E.P.F. Lausanne	EPFL	Switzerland	start	end

4. Contribution to Program / Key Action Objectives

VIBES' themes are central to the objectives of IST's User Friendly Information Society program. It has a strong and evident technical overlap with IST key action 'III. Multimedia content and tools', especially the content, representation and access action lines 'III.2. Interactive publishing, digital content and cultural heritage' and 'III.5.2 Media representation and access: new models and standards'. The ability to search and manipulate video streams at an object level will also enable more intuitive, non-technical, high-level interfaces to video, thus promoting the social goals of user-friendly-ness and IT for all citizens, and increasing access to the video component of Europe's rich cultural heritage.

Although VIBES itself will focus on enhancing human interaction with recorded video, the automatic creation and manipulation of object level video and motion representations is also likely to be a key tool for machine understanding of live video. VIBES could thus have very significant feed-through into such applications as intelligent offices and homes, intuitive non-keyboard interfaces (*e.g.* for the disabled), and automated surveillance (*e.g.* for the care of the young or elderly). VIBES also has links to the new FET 'Disappearing computer' proactive initiative. The technology it represents should be suitable for embedding in future home entertainment systems with natural interfaces, and we will be embedding computation in natural video objects, just as the disappearing computer initiative embeds it in real ones. The VIBES proof-of-concept will also contribute to the feasibility of emerging standards such as MPEG-7, and to the acceptance of existing ones like MPEG-4.

4.1 Suitability for FET Open

Despite the above-mentioned overlap with IST key action III, VIBES represents pre-competitive long-term research with a significant risk component, so it is most suitable for inclusion in the FET Open domain.

VIBES is an effort to automate the building and manipulation of very high level representations from uninstrumented video — representations sufficiently rich and persistent to support intuitive human interaction at the 'semantic' object level rather than the frame or macro-block level. These sorts of representations are available in the MPEG-4 and the advocated MPEG-7 standards, but at present they are extremely difficult to create automatically from video. Currently, content creation requires either special equipment such as 3D scanners and motion capture systems or heavy manual intervention, or restricts attention exclusively to the simplest layers of the standard (*e.g.* MPEG-2 blocks or at most flat 2D sprites, with little or no representation reuse over segments longer than a few frames, and hence little or no support for direct, natural scene-object-level interaction). In contrast, VIBES will attempt to build manipulable object and scene representations with true 3D capabilities, and object-level (rather than block or flow level) motion representations, and it will do this from natural, uninstrumented video. We believe that we can achieve this, but it is a significant challenge and well in advance of the current state of the art.

5. Innovation

VIBES has two main themes. The first is to automatically describe and index the content of video sequences or movies in a meaningful manner. The second is to automatically build 3D models of the videoed scenes and the humans moving through these scenes.

We will contrast the traditional approaches to these problem, where content analysis is often limited to the analysis of simple 2D image properties, to the VIBES approach that is rooted in the 3D scene geometry and motion.

5.1 State of the art approaches to video parsing

Existing approaches to automated parsing of a video rely mostly on segmenting the sequences into "shots," defined as continuous camera drives, that is sequences of frames filmed either from a fixed camera position, or using coherent camera motion, such as panning, rotating and zooming. These shots can then be grouped into "scenes," that is clusters of shots that present some similarity or are assumed to pertain to the same topic. It has long been recognized that these shots should be grouped into higher level semantic units and larger logical units, for example by comparing shots along larger time intervals, or using additional clues such as textual information.

Segmentation into shots is well understood. There are effective techniques that rely on histogram comparisons and use 2D image properties such as: variance of frame differences, motion continuity, variance of motion compensated frame difference, similarity between colour histograms of subsequent frames, χ^2 test on histogram difference, texture matrices, etc.

These can be work well in specific cases and can be augmented by using domain-specific knowledge. However, they suffer from some severe limitations. For example, because they capture neither the 3D nature of the scene nor its semantics content, these traditional techniques treat different views of the same scene, *e.g.* filmed by different cameras, as distinct shots that will then have to be grouped together at a higher level. Similarly, without 3D modeling, it is difficult to understand that even though a person is walking in front of the camera, thereby disturbing 2D image statistics, the sequence should nevertheless not be broken up into different shots.

5.2 The VIBES approach

The novelty of the approach we describe includes:

- We employ features with richer semantics than just intensities, colours, or textures. Rather than segmenting videos into 'shots', the project will use a multi-threaded division into 'spatio-temporal video units' (see below);
- Unlike 2D video processing, where motion segmentation and prediction is often limited to 2D flow fields, the partitioning and modelling in VIBES is based on the 3D geometry of the scene and motion (see below).

We build on recent computer vision research in three areas: advances in multiple view geometry have demonstrated that it is possible to build a three dimensional (3D) reconstruction of a static scene directly from a video sequence; advances in motion segmentation have shown that foreground and background motions can be automatically separated; advances in human tracking using parametrized kinematic models have shown that it is possible to extract human shape and dynamics from a video sequence.

We will test these approaches for the purpose of segmenting and indexing videos that show people and their surroundings. This constitutes a rich testbed for two reasons. First, people are essential in many types of videos and movies, ranging from situations comedies to newscasts to movies. Second, we live in a 3D dynamic world and full understanding of it has to go beyond 2D modeling.

5.2.1 Spatio-Temporal Video Units

These units will be characterized by the constant presence of one of the important elements one can expect to find and use for indexing purposes in a video. Such elements include:

- Objects such as faces, people, animals, cars that may be either static or dynamic and can be represented by their geometry and photometry.
- Actions that are characterized by rigid or articulated trajectories.
- Scenes that serve as backgrounds for the objects and actions, such as rooms, outdoor scenes, etc. They can be represented as layers or 2.5–D or 3D reconstructions. They are usually static, but may cause occlusions.
- Camera motions that include variations of camera external and internal parameters trajectories as well as transitions between shots.

Each one of these categories corresponds to different classes of parametric models that one might use to describe the video. A video unit can thus be understood as an instantiable model of an object, motion, location, or action, together with a bundle of parameters describing its instantiation. Segmentation then means determining the range of applicability of the model within the sequence.

This is very close to the MPEG-4 philosophy: “scenes” correspond to MPEG-4 “sprites,” that is images or layers sent once during the sequence, “objects” to “definition parameters,” that is shape and texture parameters, and “actions” to “animation parameters”.

5.2.2 The need for 3D

When a scene is filmed by either a moving camera or several different cameras, its aspect can change dramatically dues to perspective distortion and because parts of the scene come in and out of view. This is fundamentally 3D phenomenon. Furthermore, fixed scene elements must be distinguished from mobile objects. The partner’s expertise in wide-based matching and layered-representations will allow us to tell one from the other and match scenes and objects in shots from very different viewpoints.

The people also are inherently three dimensional. They can occlude each other or become visible or hidden as they move about. Understanding the actions of even one person also involves a good understanding of occlusions, as limbs and body tend to pass in front of each other.

The models and instantiation techniques developed within VIBES will be much more sophisticated than those currently used. They will not be the simple 2D filters of today. By using them to segment the image sequences, we expect to be able to provide much finer indexing capabilities than those that are now possible.

5.3 The MPEG-4 and MPEG-7 Standards

We will frequently refer to the MPEG-4 and MPEG-7 multimedia video standards in this document, so we give sketch descriptions here. See `http://drogo.cselt.stet.it/mpeg/` and `http://www.mpeg.org` for more details. Perhaps the most important point is that these standards define how to parse high-level MPEG-4 and 7 streams, *but not how to build them from existing video* — a problem that remains very difficult in practice.

MPEG-4 : The recently finalized MPEG-4 standard provides basic tools for efficient coding of multimedia scenes. Classical MPEG-1 and 2 level video coding and MPEG-3 level audio coding is provided (DCT's, motion prediction schemes), but also texture-mapped 'sprites' that can serve as mosaic-like backgrounds, VRML-like scene modelling capabilities (BIFS – BInary Format for Scenes), and detailed parametrized human head and body models that can be defined and animated at several levels of refinement. A MPEG-4 decoder requires modern, workstation-level programmable graphics and sound engines. VIBES partner EPF has developed MPEG-4 reference software and models for body animation, and donated these to ISO for distribution.

MPEG-7 : The MPEG-7 standard is still in process. It aims to provide a general, extensible multimedia content indexing framework based on XML. A number of simple shape, texture and colour descriptors have been predefined for still images, but we will aim to provide our own more advanced descriptors for object-level video hyperlinking.

6. Community Added Value and Contribution to EU Policies

6.1 European dimension of the problem

VIBES will enhance our ability to access and manipulate film and video data. This will have impact on several European-level issues:

- Economic: the film and television industries and the manufacturers of home entertainment systems are increasingly pan-European, so the economic benefits of VIBES will be equally spread throughout the partner countries, and will strengthen their combined ability to compete on the international market.
- Social: VIBES tools will enhance home entertainment and education systems by allowing easier access to existing material, and new ways to manipulate and enjoy it. The provision of more intuitive video search tools will help people of all ages and levels of education to use video more effectively and painlessly. The ability to add, remove or edit objects in the video will promote reuse and make the medium richer and more expressive.
- Heritage: improved access to Europe's rich film and video heritage will benefit education and mutual understanding, and may also help to promote tourism.
- Standards: VIBES will promote several novel MPEG-4 and MPEG-7 ideas. In particular, it is currently very difficult to parse natural video into high-level 3D entities, and this is likely to delay the widespread adoption of standards such as MPEG-4. By raising the level of abstraction at which MPEG streams can be built from natural video (from macro-block or short-lived sprite to scene object level), VIBES will encourage acceptance of these standards and promote fuller use of the facilities they offer.
- Links to other EU activities: VIBES will capitalize on results from a number of previous EU projects: VIVA, SECOND (invariants); VANGUARD, Realise, IMPACT, CUMULI (3D modelling from images); PANORAMA (virtual studio, video). It has complementarity with the following current EU projects: CAMERA (modelling of built environments from images). VIBES also has some commonality with the FET Disappearing Computer proactive initiative, in that it will allow natural human-level interaction with video, which could be built into everyday home entertainment systems.

6.2 European added value of the consortium

The work planned for VIBES will combine the specialist expertise of each of the partners, and would not be possible in any single member country, both for know-how and for budgetary reasons. For example, among other things, KTH has specialist knowledge in geometric invariants,

Weizmann in mosaicing and layered representations, INRIA in image indexing and databases, Oxford in large-scale structure from motion, Leuven in image matching and dense reconstruction, and EPFL in human modelling and the MPEG standards.

The consortium consists of experienced and well-known researchers with considerable and complementary expertise in the required areas. All of the partners have successfully contributed to cooperative projects in the past, including many European ones such as VIVA, SECOND, VANGUARD, IMPACT, PANORAMA, CUMULI, VIGOR, Improofs, CAMERA. Many of these projects and cooperations involved two or even three partners from VIBES, so the consortium has built up very good working relationships and a strong record for active, harmonious collaboration.

6.3 Contribution to EU policies

VIBES contributes to EU policies in several ways. The provision of intuitive, high-level tools for searching and manipulating video is central to the IST User Friendly Information Society key action 'III. Multimedia content and tools'. More generally, the manipulation of visual information is a central theme of the new high-tech and entertainment economies, so VIBES will contribute to the 2000-2005 EU Strategic Objective '3. A new economic and social agenda', especially 'Building a globally competitive economy based on knowledge and innovation' and 'Creating a new economic dynamism'. VIBES' intuitive, object-level video search and interaction will make Europe's rich film and video heritage more accessible and easier to use for all, regardless of age and education level. The demonstration of feasibility of the VIBES interaction techniques will contribute to the development of emerging standards such as MPEG-7, and to the acceptance of existing ones like MPEG-4. Finally, VIBES includes partners from Switzerland and Israel, thus enhancing scientific and economic cooperation and commonality of interest with these neighbouring states.

7. Contribution to Community Social Objectives

7.1 Quality of life, health and safety, working conditions

The technology developed by VIBES will contribute to quality of life directly by providing easier, faster access to and flexible object-level manipulation of our rich heritage of data stored in video format. This will enhance the video viewing experience, allowing information or content to be found more rapidly and modified if required before viewing (*e.g.* changes of viewpoint, removal of undesired objects, insertion of virtual playfellows or guides, combining objects from several videos). VIBES will enhance the entertainment and educational potential of video, and allow more intuitive object-level interfaces to it. These advances will benefit all ages, but particularly the young and old, the handicapped, and those less comfortable with complex technology.

Quite a number of interesting futuristic applications might be based on this sort of technology. For example:

Enhance your holiday videos: Build a panoramic video of the Louvre that can be viewed from any position (3D scene modelling). But remove all those other tourists who got in the way of the Mona Lisa first (object and person removal). And why not put yourself in the picture for once, instead of behind the camera (object/person insertion)? Maybe later you can buy that add-on virtual guide who will explain what all the other paintings are — after all, he can tell what you're looking at and show you other paintings by the same artist (video hyperlinking) while moving about to keep in view but out of your line of sight (virtual actor insertion). Or to really impress your friends, make him look and speak like you (animate-able human models).

To buy the guide, get on your video phone (MPEG-4 technology) and talk to one of their sales people (person tracking and resynthesis . . . or are you *sure* that nice sales girl was real?). They'll show you a number of virtual guides at work, and you can choose the one you want.

Of course, the video phone has also allowed you to work from home 3 days a week, so you save a 1 hour each way commute and have more time to play with your children. You find them playing the lead role (motion capture) in their very own video show (video resynthesis) with their favourite cartoon characters (who incidentally were programmed not only to play with them, but also to instruct them. . . you find that like you, they learn much faster that way).

Later, you are watching the football on television. To be more precise, you are *in* the match, seeing the ball from the strikers viewpoint (change of viewpoint, motion modelling). It's much more exciting that way, but also more confusing. Later maybe you'll watch yourselves star in the feature movie. . .

7.2 Employment prospects

The technology produced in VIBES will contribute to the widespread use of video in many applications and products. A few examples were given above.

7.3 Preserving the environment

Video is an important element in Europe's cultural heritage, both as a medium for preservation and transmission, and as an art form in its own right. VIBES should help to make that heritage more accessible and VIBrant for all ages. To give two examples: first, Europe has a very large repository of films, and VIBES will allow more content based searching of this; second, VIBES will enable 3D graphical models to be built from archive footage, *e.g.* of buildings that were destroyed in the war. A wider use of video may also help to promote tele-commuting, thus reducing congestion and pollution.

8. Economic Development and Scientific & Technological Prospects

8.1 Usefulness, range of applications, strategic impact

VIBES' main strategic impact will be to allow more flexible, higher level search and manipulation of video streams. This is important for those sectors already active in handling video data, like the film and TV industries, but will also be of increasing importance for any sectors that deal with human/human or human/machine interactions of some kind, *e.g.* communications, e-commerce, education, etc. Indeed, as argued before, video will become a common medium of communications also used by citizens in their day-to-day interactions.

Probable economic benefits from VIBES include:

1. The working methods of professional film and video archivers will be significantly enhanced. This will allow them to preprocess and search material more quickly, thoroughly and reliably, because queries can be formulated on a more direct, image-oriented but nevertheless high semantic level.

2. Production studios will benefit from an increased ability to build high-quality MPEG-4 streams from existing video. By encouraging more active use of the higher levels of the standard, this will allow them to improve visual quality and manipulability while reducing bandwidth, hence speeding the adoption of the standard with professionals, equipment manufacturers and consumers alike.

3. Home (or web-based) entertainment or education systems capable of object-level search of video archives and/or simple manipulations of MPEG-4 streams such as changes of viewpoint or background will become available. These will gradually be extended to allow more advanced search and manipulation.

4. Intelligent agents could put together a selection of TV programs from those that different channels offer. The selection can then be viewed at leisure. More importantly the selection can also be automatically filtered according to personal interests, based on the analysis of their content. One can even imagine that a movie can be shown with a changed, but preferred, actor in the leading role. This kind of intelligent video-on-demand systems will open new marketing opportunities – although these might be less popular. For example personalized ads can be sneaked in, tuned to the particular selection of programs.

5. More generally, as more and more products are presented through video clips over the Net, content-based video retrieval will also become an important vehicle when shopping on the Web. But there is more than efficiently finding the required data to this application. Given the kind of technology that will be developed by the project and the link with MPEG-4, shopping can subsequently also be more personalized as *e.g.* the customer herself can be shown in combination with chosen products, or their colours can be changed, or they can

be shown against another and more appropriate background like showing furniture in the customer's living room, etc. etc.

6. In education, computer-based tutoring and normal classroom teaching can be better aligned. The teacher in a video-based tutoring scheme could be replaced with the child's own teacher. This would yield more continuity from the child's perspective and, probably, more effective learning, as the same person teaching could then be seen at home (in the video-based tutoring) and at school. Alternatively the teacher in the video could be replaced by a relative.

7. In many respects, the kind of video analysis propounded here can contribute to the development of 3D TV, an area of active research by several large companies, some of which are European.

8. Just like music and text are two ways of conveying impressions and emotions, that each have lead to a series of dedicated products (music instruments, greeting cards, ...) products will emerge that allow people to edit and produce videos, where components from different videos can be combined into pieces of art or messages.

These are just some ideas, but usually the applications and possibilities of new technologies that turn out to have the largest impact are exactly those nobody foresaw. The first two in the list should occur relatively soon after the end of VIBES (given the pre-competitive, long-term research nature of the project), while the last will take somewhat longer. Nevertheless, the ability to capture and edit video quickly will become widely available. This is comparable to being able to tape music and decide on the order in which songs follow. The counterpart of composing or playing music oneself, *i.e*. manipulating video content, will certainly follow.

8.2 Exploitation and dissemination plans

Exploitation plans

VIBES represents pre-competitive long-term strategic research under the FET Open plan, and all of the partners are academic institutions rather than commercial ones, so direct commercial exploitation plans are not appropriate. Nevertheless, many of the partners have experience with spin-offs and all have regular formal and informal contacts with various companies. We have letters of support from the following companies: ArtsVideo, Eyetronics, LookThatUp, Orad and Realviz. These letters were included with the original submission.

ArtsVideo (France) is involved in interactive video construction for a number fields including advertising, training, entertainment and commercial video catalogues – in particular providing the linking for such videos. They are very interested in the tools within VIBES which will produce automatic matching of objects and scenes throughout the video.

Eyetronics (Belgium) is involved in 3D shape modeling in general and special effects for the movies and entertainment in particular. This company has won one of the European Information Technology Prizes in '98. They have expressed a keen interest in the 3D modeling activities of both scenes and actors. They also have an interest in 3D related e-commerce. Hence, the tools to produce 3D models directly out of video sequences is very close to their core business.

LookThatUp (France) are a visual e-commerce enabler by linking any digital image or video sequence to merchant sites. The ability to effectively index video sequences is an essential enabling technology to this company. They are also interested in 3D modeling and representation of scenes and objects.

Orad (Israel) have expressed an interest in the hyperlinking and classification results of the project. This company is very active in the area of virtual studios, where they offer state-of-the-art technology in terms of the flexibility in relative camera and object motions and positions. Being able to analyze scenes and to disentangle objects and backgrounds is an important asset in this area. Also, Orad is interested in the analysis of sport scenes. Moreover, they have an interest in enhanced video browsing capabilities over the Internet.

Realviz (France) are active in special effects for the movies as well. They provide software tools that dramatically reduce the time required to produce high-quality computer-generated images, animation and 3D models. For instance, they already offer 3D modeling services based on pure image data. For this company it is interesting to see how the level of automation of such reconstructions could be increased further, and how robustness against occlusions could be enhanced. It is also relevant to extract actors from scenes and to model their motions. They are also very interested about new advances in the fields of virtual reality and augmented reality environments.

As this list shows, the partners have direct contacts with a number of high-tech companies active in the 'new economy', each of which is interested in the project and well-placed to bring exploitable results quickly to the market. For Europe as a whole, it is important that such companies thrive, as the old continent's position in these new areas is weak compared to that of the US.

Wider Dissemination

Scientific results will of course be disseminated through the usual scientific channels (journals, international conferences and workshops).

The consortium considered the possibility of forging formal links with MPEG-4 and MPEG-7 standards community, but decided that this would not be appropriate given the pre-competitive, investigative nature of the project. Thematically, VIBES is very strongly aligned with MPEG-4 and MPEG-7. However tying our low level implementations to these standards would commit us to a great deal of detailed professional implementation work, thus wasting valuable research time. We feel that standards-conformance is more appropriate for an industrial reimplementation than for an investigative research project. Notwithstanding this, good informal contact with the standards community will be assured by EPFL, which is an active contributor to several standards working groups including MPEG-4 and MPEG-7.

Apart from the directly interested industrial contacts listed above, the consortium will actively maintain a web site for wider dissemination. The site will contain tutorials, on-line demos, and also original and modified videos to demonstrate the results through VIBES' own natural medium. We considered several other dissemination channels such as special-purpose industrial workshops, but (like many other organizations) we feel that these are not cost- and time-effective compared to web publicity. Our experiences with past European projects suggest that very few people attend such workshops owing to high time and travel costs, other commitments, etc. A web demo is much more flexible: it is available 100% of the time from anywhere by anyone. It is viewable at

whatever level of detail is desired. It can be updated throughout the project, made interactive to give hands-on experience, perhaps run on your own data, etc. Hence, we decided to make the web site central to our wider dissemination strategy.

However, near the end of the project, the consortium will organize an academic workshop associated with a major computer vision conference, *e.g.* as a sequel to the successful 'SMILE' series of workshops. The VIBES work will be presented at this workshop.

9. Workplan

9.1 General Description

VIBES is a 3 year FET Open R&D project. The workplan contains two interconnecting threads: video segmentation, matching and classification leading to a video hyperlinking demonstrator; and scene and human modelling leading to a video resynthesis demonstrator.

1. Segmentation, matching and classification: This thread will automatically parse the input video into primitive entities (spatio-temporal video units) including objects, scenes and actions. These will be matched across images, classified, and condensed into forms suitable for element searches and video hyperlinking. The hyperlinked video will provide indexing capabilities analogous to MPEG-7, supporting operations such as element search and video editing and manipulation.

2. Scene and human modelling: This thread will produce 3D texture mapped scene models and animate-able 3D models of humans from the segmented natural video streams. The output will be suitable for use in VRML-2 or MPEG-4 like video synthesis modules, and the demonstrator will show the human models we have built moving around in the created scene model.

Work organization: VIBES is organized into ten workpackages (WP's) covering three major classes of activity: research and development of algorithms (5 WP's); integration of scientific results into technology demonstrators (2 WP's); and Management, Assessment & Dissemination

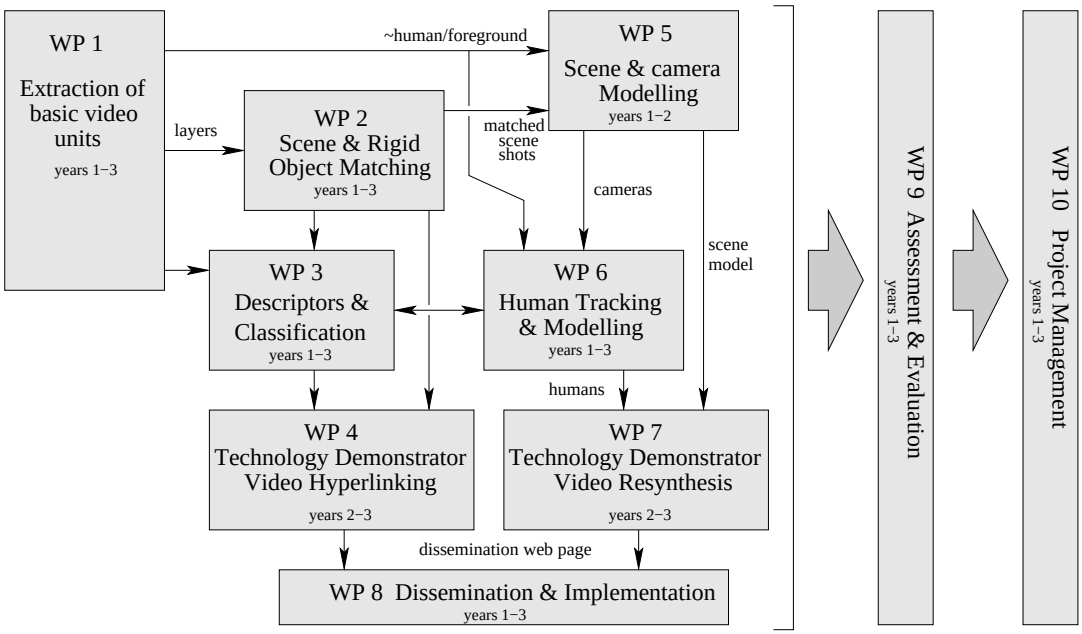

Figure 9.1: Interconnections among the workpackages.

(3 WP's). Each WP is described in detail in the following sections. Each WP runs for either 2 or 3 years and is subdivided into 12 month work-tasks ending in milestones and deliverables. The goals become progressively more sophisticated and automated as time progresses. The main interactions among the workpackages are sketched in figure 9.1.

To understand the functionality of the scientific workpackages, consider the following example. Suppose a video begins with a shot of a person walking down a corridor and turning into a room, and then a shot from a different camera inside the room. WP 1: *Extraction of basic video units* will compute two overlapping representations of this: a traditional partition into two shots ('shot' video units), and a single 'person' video unit covering the person in both shots. These representations will be passed to the other WP's. WP 2: *Scene and Rigid Object Matching* will determine if the same person or the same corridor or room appears again later in the video, and if so record the information for hyperlinking. WP 3: *Descriptors and Classification* will identify the 'person' video unit as a person walking. This will also be recorded for a searchable key. WP 6: *Human Tracking and Modelling* will track and model the person, and WP 5: *Scene and Camera Modelling* will build a 3D model of the room and corridor using the original sequence together with any other frames in which the room/corridor appear, as determined by WP 2. Finally, WP 4: *Video Hyperlinking* will provide a graphical user interface so that all the frames in which the person appears, or in which there is any person walking, can be viewed; and WP 7: *Video Resynthesis* will enable the person and/or the scene to be changed for another one.

Results will be transferred between partners in the form of data files and code. Examples of data files include: a list of video units and for each item a list of frames in which it appears; camera matrices for each frame of a sequence; a VRML model of a scene. As several of the partners are already using a common software system (the TARGETJR/VXL C++ libraries), we do not foresee any problem with the code transfer path which has worked well in previous projects.

9.2 Workpackage List

Work-package No.	Workpackage Title	Lead Con-tractor	Person months	Start-End month	Phase	Deliv-erable No.
1	Extraction of Basic Video Units	WIS	46	1–36		D 1.*
2	Scene and Rigid Object Matching	OXF	93	1–36		D 2.*
3	Descriptors and Classification	KTH	76	1–36		D 3.*
4	Technology Demonstrator: Video Hyperlinking	INR	17	13–36		D 4.*
5	Scene and Camera Modelling	KUL	30	1–24		D 5.*
6	Human Tracking and Modelling	EPF	66	1–36		D 6.*
7	Technology Demonstrator: Video Resynthesis	INR	19	13–36		D 7.*
8	Dissemination and Implementation	KTH	12	1–36		D 8.*
9	Assessment and Evaluation	KTH	10	1–36		D 9.*
10	Project Management	KTH	12	1–36		D 10.*
	TOTAL		**381**			

9.3 Workpackage Descriptions

See following pages.

WP 1. Extraction of Basic Video Units

WP Number:	1	Leader + Participants:	WIS	KTH	INR			
Start–End Month:	1–36	Person-months:	34	6	6			

Objectives

Construct a video front-end to segment natural video streams into "spatio-temporal video units": contiguous 2D video regions defined across multiple frames, representing static backgrounds, moving objects or people. Includes cut detection, extraction of key frames, and object-based and action-class based segmentation methods.

This is a generalization of the camera cut detection routinely used to segment video into shots. The methods employed here are primarily based on motion — of the camera and of independently moving objects and people.

Description of work

Year 1. WP 1.1 Simple shot detection. Register backgrounds for a panning camera. Use residual motion to segment simple independently moving rigid objects.

Year 2. WP 1.2: More sophisticated segmentation. Segment independently moving rigid and non-rigid objects. The methods used here will be based on analysis of spatio-temporal brightness volumes (as opposed to segmenting the spatial and temporal dimensions independently). This will allow segmentation of more complex sequences.

Year 3. WP 1.3 Action based segmentation. Segmentation into action classes, *e.g.* segment sequence of a human in motion into standing/walking/running.

Deliverables

Month 12. D 1.1 Demonstrator: Simple temporal segmentation into shots and simple spatial segmentation into small independently moving rigid objects.

Month 24. D 1.2a Demonstrator: Spatio-temporal video segmentation, based on the analysis of brightness volumes.

D 1.2b Paper: Spatio-temporal video segmentation.

Month 36. D 1.3a Demonstrator: Segmentation based on action classes.

D 1.3b Paper: Advanced segmentation methods.

Milestones and expected results

Month 3. M 1.1a Preliminary selection of test video suite to be used by all partners.

Month 6. M 1.1b Rough shot segmentation of selected test videos (ASCII files listing cut frames for use in other WP's).

Month 12. M 1.1c Improved shot segmentations and independently moving rigid objects for panning sequences of test videos.

Month 24. M 1.2 Segmentation of test videos based on analysis of spatio-temporal brightness volumes.

Month 36. M 1.3 Segmentations of test videos into action classes.

Input from other WPs
None.

Output to other WP's

WP 2: provides shots, key frames, and objects to be matched.

WP 3: isolates objects and humans for description and classification.

WP 5: moving object delineation to enable removal for background modelling.

WP 6: isolates humans as means of initializing tracking.

Fall back position and assessment of difficulty
There is little risk in WP1.1. However, manual segmentations can be provided if automation does fail. For WP1.2 the fall back position is rigid objects rather than non-rigid objects. WP1.3 is the most ambitious. It will be tested first on scenes viewed with static cameras, then on scenes with panning cameras. If action based classification does fail here, then there is an alternative approach available in WP6 based on hidden state models.

Evaluation criteria
The test suite will be hand segmented into each of the spatio-temporal units: contiguous sequences of frames based on: camera motion; object motion; and human action. The measure of success is the similarity of the automated and hand based segmentations.

WP 2. Scene and Rigid Object Matching

WP Number:	2	Leader + Participants:	OXF	KUL	WIS	INR		
Start–End Month:	1–36	Person-months:	24	31	24	14		

Objectives

Identify the same rigid object or 3D scene in different shots, and find coarse geometric correspondences. Camera viewpoints and image scales may differ widely. Object appearance may change due to different viewing direction, lighting, weather. Matching should be efficient, even with many candidate shots.

The methods of this WP are based on the image appearance and geometry, together with multi-view relations which arise from rigid object and scene geometry.

Description of work

Year 1. WP 2.1 Methods for pairwise coarse wide-baseline matching. Interest point and affine invariant-neighbourhood detectors and descriptors, simple inefficient algorithms. The fundamental matrix is the multi-view constraint.

Year 2. WP 2.2 Strategies for efficient matching (invariants, indexing), 3 image matching. Richer descriptions. Use of the trifocal tensor.

Year 3. WP 2.3 Extend to multi-image wide-baseline matching with verification. Improve reliability and efficiency. Sequence-to-sequence alignment of parallel video streams.

Deliverables

Month 12. D 2.1a Demonstrator: Simple wide-baseline matching. Pairwise images, query specified by hand.

D 2.1b Paper: Wide baseline matching.

Month 24. D 2.2a Demonstrator: Efficient inter-shot matching. Invariant based indexing, query defined by WP1 segmentation method.

D 2.2b Paper: Efficient inter-shot matching.

Month 36. D 2.3a Demonstrator: Multi-image matching. Sequence-to-sequence alignment of parallel video streams. Finding the same object/scene in all shots of a video using combined matching techniques.

D 2.3b Papers: Multi-image matching.

Milestones and expected results

Month 12. M 2.1 Pairwise wide-baseline matches of test videos.

Month 24. M 2.2 Invariant-based matches of video units (objects/scenes) from test videos.

Month 36. M 2.3 Robust whole video matches of video units from test videos.

Input from other WP's

WP 2.1: Shots and key frames from WP 1.1

WP 2.2: Objects detected in WP 1.1/1.2.

WP 2.3: Same inputs as WP 2.2.

Output to other WP's

WP 4: The object/scene matching provides one of the indexing fields that is used in the Hyperlinking GUI.

WP 5: The matched shots of scenes provides additional views to be used in building scene models.

Fall back position and assessment of difficulty

The appearance of objects often varies significantly with viewpoint, so pose independent matching is an ambitious goal. The fall back position is a restriction on pose differences, *e.g*. smaller than 30 degrees compared to the frame in which the user has delineated the object, and with the object covering at least 20% of the image. This is still beyond the typical rotations that traditional methods can tolerate. Similarly the illumination changes between frames can be restricted. Other than these restrictions, this WP is low risk. However, the challenge is in making the algorithms efficient.

Evaluation criteria

The measure of success for matching is the percentage of detections of rigid, well-textured objects/scenes visible elsewhere in a video, with respect to the total number of shots (not frames) in which they appear. Results will be reported in terms of false positives and false negatives.

WP 3. Descriptors and Classification

WP Number:	3	Leader + Participants:	KTH	WIS	INR	KUL	OXF	
Start–End Month:	1–36	Person-months:	24	29	15	4	4	

Objectives

Characterize or classify the content of natural video frames or shots, for use in indexing and matching/hyperlinking. We will start by considering relatively simple descriptors for particular objects and actions. Later, we will work towards more generic classification into high-level object categories. The results will be used for object-level matching and hyperlinking. The objects we are particularly interested in are humans and their actions.

Description of work

Year 1. WP 3.1 Simple descriptors. Stationary cameras. Face detection from individual images with temporal verification. Recognition of same face in later shots. Detection of humans using prototype matching.

Year 2. WP 3.2 Richer descriptions, more complicated shots. Recognition of simple human activities (walking) from 2D motions. Appearance based classification of simple objects.

Year 3. WP 3.3 Move towards more ambitious generic classification. Classification based on local texture for simple object types including: humans, cars, and animals. Person ID from gait. Scene classification including indoors/outdoors; natural/man-made; closed space/open space. Appearance based classification in cluttered scenes.

Deliverables

Month 12. D 3.1 Demonstrator: Simple descriptors. Simple face recognition. Detection of humans using prototype matching.

Month 24. D 3.2a Demonstrator: Rich descriptors. Recognition of human activities, appearance based recognition.

D 3.2b Paper: Recognition of faces or simple objects.

Month 36. D 3.3a Demonstrator: Generic classification. Qualitative scene categorization. Object classification.

D 3.3b Paper: Generic classification.

Milestones and expected result

Month 12. M 3.1 Simple descriptors for test videos, faces, humans from motion.

Month 24. M 3.2 Richer descriptors from test videos, activities, object recognition.

Month 36. M 3.3 Generic classifications from test videos.

Input from other WP's

WP 3.2: Objects detected in WP 1.1–1.2;

WP 3.3: Shots and objects detected in WP 1.1–1.2;

Output to other WP's

WP 4: The semantic scene content provides indexes/keys suitable for matching and hyperlinking.

WP 6: Moving objects classified as humans provide an initialization for the human tracking.

Fall back position and assessment of difficulty

The fall back position of face recognition (*i.e*. recognizing the same person in various shots) is general face detection, which is low risk.

Restrictions may be placed on pose changes and scale changes for object classification. Classification into object types using texture is more likely to be successful for animals and cars than humans.

Generic scene classification is ambitious. If this is unsuccessful the main loss is fewer categories and conjunctions for indexing and search in WP 4.

Evaluation criteria

The measure of success is the proportion of correctly classified shots, *i.e*. false positives and false negatives for object and scene classification.

WP 4. Technology Demonstrator: Video Hyperlinking

WP Number:	4	Leader + Participants:	INR	WIS	KUL	OXF	KTH	
Start–End Month:	13–36	Person-months:	4	4	4	4	1	

Objectives

Provide a public web-based demonstrator to showcase VIBES' video-hyperlinking technology. This will be a lightweight web GUI accessing a set of programs from WP 1,2,3. It will allow the user to 'surf' through a number of video clips, by clicking on objects, scenes, actions, etc. This demonstrates VIBES' high-level MPEG-7 like indexing capabilities, extracted from natural video.

Description of work

Year 2. WP 4.2 Initial version of the GUI. Simple hyperlinking of video objects extracted and matched in WP 1,2.

Year 3. WP 4.3 More sophisticated GUI. Hyperlinking by generic category and actions, using WP 3. If resources permit: also demonstrate searching on conjunctions of events (*e.g.* two people meeting), hyperlink-based editing/manipulation.

Deliverables

Month 24. D 4.2 Demonstrator: Initial video hyperlinking GUI. Simple hyperlinking, serves only static key frames.

Month 36. D 4.3 Demonstrator + White-paper: Final video hyperlinking GUI. Richer indexing using year 3 results, more test videos, build-your-own index (resources permitting). White paper describing techniques used in demonstrator.

Milestones and expected results

Month 24. M 4.2 Initial interactive video hyperlinking demo on the web.

Month 36. M 4.3 Final interactive video hyperlinking demo on the web, plus white-paper.

Input from other WP's

WP 4.2: matches on scene and object shots from WP 2.

WP 4.3: classification of objects, human actions, and scenes from WP 3.

Output to other WP's

None.

Fall back position and assessment of difficulty

This is a display and search tool for WP's 2 and 3 with well defined inputs (video sequences, ASCII text files) and as such there is a low research risk.

Evaluation criteria

The measure of success is more qualitative here: the speed and ease of use of the GUI.

WP 5. Scene and Camera Modelling

WP Number:	5	Leader + Participants:	KUL	OXF	INR			
Start–End Month:	1–24	Person-months:	12	12	6			

Objectives

Build texture-mapped Euclidean 3D VRML scene models from natural monocular video sequences. Also compute accurate, low-jitter camera motions and calibrations for use by WP 6. The model will initially be built from a single shot containing a tracking camera. In year 2 multiple shots will be combined to extend and strengthen the model, using matching from WP 2.

Description of work

Year 1. WP 5.1 Scene modelling (geometry and texture) and camera pose from single video shots. Assumes moving camera, static scene. Cameras are output for use in WP 6.

Year 2. WP 5.2 Extend method to combine multiple shots (for improved coverage, more accurate models), and ignore independently moving objects. This will use WP 2 wide-baseline matching and WP 1 moving object segmentation. Also include levels of detail for texture.

Deliverables

Month 12. D 5.1 Demonstrator: VRML models and camera motion. Built from one-shot test videos (*e.g*. classic Hollywood scenes).

Month 24. D 5.2a Demonstrator: Multi-shot modelling. VRML models and camera motion built from multiple shots. Classic film scenes (background after foreground objects removed).

D 5.2b Paper: Video-based scene modelling.

Milestones and expected results

Month 3. M 5.1a Initial estimation of camera motions in test videos (*e.g*. tracking shots from classic films), for use in WP 6 development.

Month 12. M 5.1b Scene and camera reconstructions extracted automatically from test videos (single tracking shots).

Month 24. M 5.2 Automatic scene and camera reconstructions of test videos, combining multiple shots.

Input from other WP's

WP 1: Video of static scene with regions containing independently moving objects labeled for excision.

WP 2: Wide baseline correspondence methods.

Output to other WP's

WP 4: camera motions, for video indexing by camera motion.

WP 6: camera position in the scene, for use in human modelling.

WP 7: scene models for technology demonstrator.

Fall back position and assessment of difficulty
This WP builds on strong results obtained in previous European projects involving the partners, notably VANGUARD. These results will be further automated and specialized to video streams, and improved wide-baseline matching results from WP 2.1 will be integrated. The consortium has considerable experience in this field, so there is little risk of failure. If necessary we could even choose test videos with scenes simple enough to be reconstructed using existing methods. For example, restrict shots to those where the independently moving objects occupy less than 25% of the image, and where depth discontinuities don't cause order reversals between corresponding points in successive frames.

Evaluation criteria
VRML models of static scenes and camera positions, built from uncalibrated natural video sequences. Ultimately the best measure of success is verisimilitude: can the 3D models be used to produce realistically looking scenes where new objects have been included or original ones have been removed. This calls for precision in depth and geometry.

WP 6. Human Tracking and Modelling

WP Number:	6	Leader + Participants:	EPF	INR	KTH			
Start–End Month:	1–36	Person-months:	36	18	12			

Objectives

Investigate how far we can go towards automatic tracking of articulated human motions, and extraction of complete, animation-ready 3-D models of people from natural uninstrumented monocular video sequences. The models should capture (i) 3D skeleton and flesh geometry, (ii) appearance, and (iii) 3D motion characteristics, in a compact form suitable for producing convincing synthetic views of the moving person. The input views may be head-only, upper body, or full body shots.

Description of work

Year 1. WP 6.1 Simple 3D modelling of the upper body of a static human from a moving camera. Simple articulated body tracking (clean backgrounds, simple motions, hand initialization).

Month 24. WP 6.2 High-resolution 3D modelling of a complete static human body from a moving camera. More advanced articulated body tracking (some clutter, more complex motions, self occlusion, automatic initialization).

Year 3. WP 6.3 Combine articulated tracking and 3D human modelling to handle a moving camera and a moving subject. Improve robustness. Merge models from several different shots.

Deliverables

Month 12. D 6.1a Demonstrator: Simple human modelling and tracking. 3D upper-body modelling from a moving camera. Simple hand-initialized articulated body tracking.

D 6.1b Paper: 3D body modelling.

Month 24. D 6.2a Demonstrator: Advanced human modelling and tracking. Full high-resolution 3D body modelling from a moving camera. More advanced articulated body tracking (clutter, complex motions, automatic initialization).

D 6.2b Papers: (i) 3D body modelling; (ii) articulated tracking.

Month 36. D 6.3a Demonstrator: Combined human modelling and tracking. Articulated body tracking and 3D human modelling from a moving camera. Merging of partial reconstructions across shots.

D 6.3b Paper: Combining human tracking and modelling.

Milestones and expected results

Month 12. M 6.1 Simple upper-body models and articulated tracks from test videos.

Month 24. M 6.2 More detailed full-body models and more complex motion tracks from test videos.

Month 36. M 6.3 Moving body models from test videos, using combined tracking and body modelling.

Input from other WP's

WP 1: video windows on moving people.

WP 5: 3D camera motions.

Output to other WP's

WP 7: parametric articulation + appearance + motion models of humans for video resynthesis demonstrator.

Fall back position and assessment of difficulty

This is an ambitious, investigative WP with relatively high risk, consistent with the nature of a FET Open research project. However, there is considerable scope for raising or lowering the final goals in response to our progress during the project. The central goal of automated human tracking and modelling will not change, but we will select more or less difficult video sequences according to our current capabilities: varying the amount of background clutter and lighting variation, amount of prior information required, speed and complexity of motion, type of clothing, number and overlap of people, level of modelling detail, occlusions. In this way, the WP should not fail completely, even if its goals have to be restricted.

There is little downstream risk to the project from this WP, as the only WP that depends directly on it is the demonstrator WP 7, which can be adapted to show whatever we achieve in this WP.

Evaluation criteria

The evaluation criterion is simply the extent to which we can automatically build visually realistic motion and appearance models of people from uncalibrated video sequences. Our current goal for the end of the project is to be able to do this for a person moving around naturally but simply in front of a camera.

WP 7. Technology Demonstrator: Video Resynthesis

WP Number:	7	Leader + Participants:	INR	EPF	KUL	OXF		
Start–End Month:	13–36	Person-months:	5	7	4	3		

Objectives

Provide a public web-based demonstrator to showcase VIBES' video resynthesis technology. This will be a lightweight web GUI accessing a set of programs from WP 1,2,5,6. It will allow the user to move around in the reconstructed scene, and perform simple manipulations of reconstructed human model(s) in the scene (*e.g.* animation, replacement with another model). This demonstrates VIBES' high-level MPEG-4 like model-building from natural video.

Description of work

Year 2. WP 7.2 Simple changes of scene elements (*e.g.* viewpoint). Insertion of novel static objects and static human models in scene. Simple manipulations of human models (move arms).

Year 3. WP 7.3 More advanced insertion: moving human and camera, occlusions. View results of human replacement. GUI to provide web access to this demonstrator.

Deliverables

Month 24. D 7.2 Demonstrator: Simple changes of scene elements (*e.g.* viewpoint). Simple manipulations of a reconstructed human model (*e.g.* apply motions). Initial insertion of human model in reconstructed scene.

Month 36. D 7.3 Demonstrator + white-paper: WWW video resynthesis GUI for interactive access. More advanced insertion: moving human and camera, occlusions. View results of human replacement. White paper describing techniques used in demonstrator.

Milestones and expected results

Month 24. M 7.2 Initial video resynthesis demo on the web.

Month 36. M 7.3 Final interactive video resynthesis demo on the web.

Input from other WP's

WP 5: Scene models and camera poses.

WP 6: Human motion and appearance models.

Output to other WP's

None.

Fall back position and assessment of difficulty

This WP is simply a demonstrator for the technology developed in WP 5 and 6. In itself it has relatively low risk as it only requires writing a GUI and some glue code. The main risk is from the upstream WP's 5,6. If either of these fail to achieve their full goals, the demonstrator will be downgraded to show what is available.

Evaluation criteria
A working WWW demonstrator of the technology available from WP 5 and 6.

WP 8. Dissemination and Implementation

WP Number:	8	Leader + Participants:	KTH	WIS	INR	KUL	OXF	EPF
Start–End Month:	1–36	Person-months:	7	1	1	1	1	1

Objectives

Workpackage for efforts devoted to disseminating the results of VIBES. Dissemination will be assured by: the project web pages (descriptions of results, WWW demonstrators from WP's 4 and 7); a VIBES-oriented workshop in year 3; scientific publications; informal contacts with companies who have expressed interest in the results (see letters of interest); and informal contacts with the MPEG-4 and MPEG-7 standards efforts.

Description of work

Month 1–36. WP 8.1 Create initial WWW Project Presentation by month 6, then update the WWW pages continuously throughout the project as new results and scientific papers are generated. This activity will be monitored by the Coordinator's assistant at six-monthly intervals.

Month 1–6. WP 8.2 Create Dissemination and Use plan.

Month 25–36. WP 8.3 The consortium will organize an open scientific workshop on VIBES themes, at a major computer vision conference, and prepare the VIBES Technology Implementation plan.

Deliverables

Month 6. D8.1 Initial WWW Project Presentation.

D8.2 Dissemination and Use plan.

Month 36. D8.3 Technology Implementation plan.

The proceedings of the scientific workshop at the end of the project will also be published, *e.g.* in the Springer-Verlag LNCS series.

Milestones and expected results

Month 6. M 8.1 Initial WWW Project Presentation D 8.1 available.

M 8.2 Dissemination and Use plan D 8.2 available.

Month 36. M 8.3 Scientific workshop held on VIBES themes.

Technology Implementation plan D 8.3 available.

WP 9. Assessment and Evaluation

WP Number:	9	Leader + Participants:	KTH	WIS	INR	KUL	OXF	EPF
Start–End Month:	1–36	Person-months:	5	1	1	1	1	1

Objectives

Monitor each Work Package according to the self-evaluation and fall-back criteria specified in its WP description, and if necessary recommend modifications to its workplan to achieve the best attainable goals for the overall project.

Description of work

Month 6,12,18,24,30,36. Each WP leader will assess progress on his WP against its evaluation criteria, and report to the project coordinator who will assess the overall progress and conformance with the workplan. Remedial actions will be recommended as necessary, for implementation by WP 10, Project Management.

Deliverables

Month 12,24,36. (Deliverables D9.1–9.3). Short 1 page executive updates summarizing progress so far, project status, advances or delays, *etc*, and also giving plans for remedial actions to correct any problems that may arise.

Milestones and expected results

Month 12,24,36. Delivery of progress updates D9.1–9.3.

WP 10. Project Management

WP Number:	10	Leader + Participants:	KTH	WIS	INR	KUL	OXF	EPF
Start–End Month:	1–36	Person-months:	7	1	1	1	1	1

Objectives

Manage the project: project management committees, annual reports and reviews. Implement and monitor any changes of workplan recommended by WP 9.

Description of work

Month 6,12,18,24,30,36. 6 monthly project management committees, including discussion of latest

Month 12,24,36. Annual reports and reviews.

Deliverables

Month 12,24,36. (D 10.1–10.3). Annual periodic progress reports.

Milestones and expected results

Month 12,24,36. Annual reports available.

9.4 Deliverables List

Deliv. No.	Deliverable Name	WP no.	Est. person months	Del. Type	Security level	Month Due
1.1	Segmentation into basic video units	1.1	15	demo	pub.	12
1.2a	Spatio-temporal video segmentation	1.2	8	demo	pub.	24
1.2b	Spatio-temporal video segmentation	1.2	7	paper	pub.	24
1.3a	Segmenting complex scenes, action classes	1.3	8	demo	pub.	36
1.3b	Advanced segmentation methods	1.3	8	paper	pub.	36
2.1a	Simple wide-baseline matching	2.1	16	demo	pub.	12
2.1b	Wide baseline matching	2.1	15	paper	pub.	12
2.2a	Efficient inter-shot matching	2.2	16	demo	pub.	24
2.2b	Efficient inter-shot matching	2.2	15	paper	pub.	24
2.3a	Multi-image matching	2.3	16	demo	pub.	36
2.3b	Multi-image matching	2.3	15	paper	pub.	36
3.1	Simple descriptors	3.1	25	demo	pub.	12
3.2a	Rich descriptors	3.2	13	demo	pub.	24
3.2b	Recognition of faces or simple objects	3.2	12	paper	pub.	24
3.3a	Generic classification	3.3	13	demo	pub.	36
3.3b	Generic classification	3.3	13	paper	pub.	36
4.2	Initial video hyperlinking GUI	4.2	8	demo	pub.	24
4.3	Final video hyperlinking GUI	4.3	9	demo+report	pub.	36
5.1	VRML models and camera motion	5.1	15	demo	pub.	12
5.2a	Multi-shot modelling	5.2	8	demo	pub.	24
5.2b	Video-based scene modelling	5.2	7	paper	pub.	24
6.1a	Simple human modelling and tracking	6.1	11	demo	pub.	12
6.1b	3D body modelling	6.1	11	paper	pub.	12
6.2a	Advanced human modelling and tracking	6.2	11	demo	pub.	24
6.2b	3D body modelling + articulated tracking	6.2	11	papers	pub.	24
6.3a	Combined human modelling and tracking	6.3	11	demo	pub.	36
6.3b	Combining human tracking and modelling	6.3	11	paper	pub.	36
7.2	Initial video resynthesis	7.2	9	demo	pub.	24
7.3	Video resynthesis, GUI	7.3	10	demo+report	pub.	36
8.1	Initial WWW Project Presentation	8.1	4	web pages	pub.	6
8.2	Dissemination and Use Plan	8.2	3	report	pub.	6
8.3	Technology Implementation Plan	8.3	5	report	pub.	36
9.1	Assessment & Evaluation Summary	9.1	3	summary	pub.	12
9.2	Assessment & Evaluation Summary	9.2	3	summary	pub.	24
9.3	Assessment & Evaluation Summary	9.3	4	summary	pub.	36
10.1	Annual Report	10.1	4	report	pub.	12
10.2	Annual Report	10.2	4	report	pub.	24
10.3	Annual Report	10.3	4	report	pub.	36

9.5 Project Planning and Timetable

	Gannt Chart showing the relative timing of the workpackages												
		Start Month / End Month											
WP no.	**Workpackage**	**1 3**	**4 6**	**7 9**	**10 12**	**13 15**	**16 18**	**19 21**	**22 24**	**25 27**	**28 30**	**31 33**	**34 36**
1	**Extraction of basic video units**	■	■	■	■	■	■	■	■	■	■	■	■
1.1	Shot detection and independent motion	■	■	■	■								
1.2	Rigid and non-rigid objects					■	■	■	■				
1.3	Action based segmentation									■	■	■	■
2	**Scene and rigid object matching**	■	■	■	■	■	■	■	■	■	■	■	■
2.1	Pairwise matching	■	■	■	■								
2.2	Improved efficiency					■	■	■	■				
2.3	Sequence-wide matching									■	■	■	■
3	**Descriptors and classification**	■	■	■	■	■	■	■	■	■	■	■	■
3.1	Simple objects and actions	■	■	■	■								
3.2	Richer descriptors					■	■	■	■				
3.3	More generic classes of objects and actions									■	■	■	■
4	**Technology demonstrator: video hyperlinking**					■	■	■	■	■	■	■	■
4.2	Basic hyperlinking					■	■	■	■				
4.3	More sophisticated hyperlinking									■	■	■	■
5	**Scene and camera modelling**	■	■	■	■	■	■	■	■				
5.1	Simple shots and static scenes	■	■	■	■								
5.2	Multiple shots and motions within shots					■	■	■	■				
6	**Human tracking and modelling**	■	■	■	■	■	■	■	■	■	■	■	■
6.1	Simple modelling and tracking	■	■	■	■								
6.2	Accurate modelling and robust tracking					■	■	■	■				
6.3	Combined human modelling and tracking									■	■	■	■
7	**Technology demonstrator: video resynthesis**					■	■	■	■	■	■	■	■
7.2	Simple actor insertion and deletion					■	■	■	■				
7.3	Actor replacement									■	■	■	■
8	**Dissemination and implementation**	■	■	■	■	■	■	■	■	■	■	■	■
8.1	WWW Project presentation + updates	■	■		■		■		■		■		■
8.2	Dissemination and use plan	■	■										
8.3	Technology implementation plan											■	■
9	**Assessment and evaluation**	■	■	■	■	■	■	■	■	■	■	■	■
9.1	Year 1 assessment and evaluation		■		■								
9.2	Year 2 assessment and evaluation						■		■				
9.3	Year 3 assessment and evaluation										■		■
10	**Project management**	■	■	■	■	■	■	■	■	■	■	■	■
10.1	Year 1 management	■	■	■	■								
10.2	Year 2 management					■	■	■	■				
10.3	Year 3 management									■	■	■	■

9.6 Graphical Presentation of Project Components

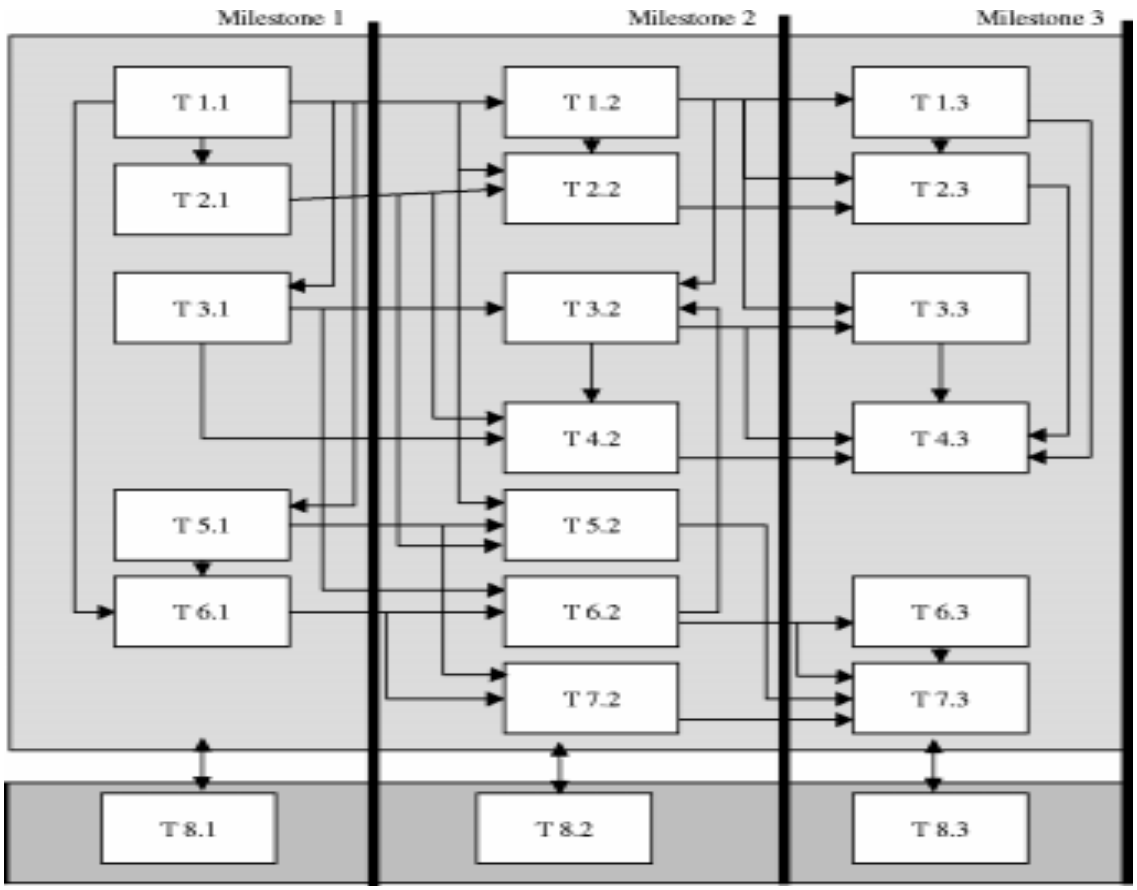

Figure 9.2: PERT diagram showing interconnections among the scientific workpackages. The Assessment and Management workpackages WP 9,10 monitor all of the scientific ones, and are not shown for clarity.

9.7 Project Management

Workpackages 8–10 are dedicated to project management and dissemination of results. All partners will contribute to these efforts, which will be lead by the project coordinator. As requested by the Commission, we have broken these efforts down into:

- **WP 8 Dissemination and Implementation:** This WP covers the dissemination of VIBES results including development and regular updating of the WWW pages, workshops, and the Dissemination and Use plan (month 6) and Technology Implementation plan (month 36).

- **WP 9 Assessment and Evaluation:** This WP assesses the project status and conformance to the workplan at regular 6 monthly intervals, and recommends workplan corrections to keep the whole project on course. See §9.7.2 for more details.

- **WP 10 Project Management:** This WP covers project management committee activity, communication flow and adoption and implementation of the recommendations of the Assessment and Evaluation WP.

To distribute the burden of project management, each scientific workpackage has been assigned a **Workpackage Leader** as follows:

	Overall project coordination	KTH
WP1	Video segmentation	Weizmann
WP2	Scene & object matching	Oxford
WP3	Descriptors & classification	KTH
WP4	Video hyperlinking demonstrator	INRIA
WP5	Scene & camera modelling	Leuven
WP6	Human modelling	EPFL
WP7	Video resynthesis demonstrator	INRIA
WP8	Dissemination & implementation	KTH

The Workpackage Leaders will be responsible for coordinating the scientific and development work within "their" work package, including:
— Keeping the Project Coordinator informed about progress as defined to WP 9, including any delays or problems that may arise.
— Assisting the Project Coordinator with the preparation of reports and deliverables relating to their work package, *e.g.* writing task summaries.
— Presenting a synthesis of the results and status of their package at review meetings, highlighting collaborations and pointing out possible weaknesses.

A part-time **Coordination Assistant** at the coordinating site will be responsible for: the practical organization of meetings; interface with the commission; collecting, distributing and maintaining a database of all deliverables and reports; maintaining the VIBES WWW pages and related dissemination efforts; and fielding any inquiries about the project.

9.7.1 Decision making and communication flow

Project management decisions will be made by the Project Coordinator in consultation with a Project Management Committee formed from the Workpackage Leaders (1 vote per site). The management committee will normally meet every 6 months, but exceptional meetings may be called by any partner if required. These exceptional meetings will by default be held by telephone to save time and cost. If any serious conflicts should arise, these will be resolved by a full meeting of the Project Management Committee, with the Project Coordinator having the casting vote in the event of a tie.

Meetings between the partners will be held every six months with a kick-off meeting at the start of the project. These will be informal scientific meetings where the partners present work in progress and plan future efforts. Other parties interested in the project (the Commission, the reviewers, industrial associates, ...) are welcome to attend these informal meetings if they wish to. Each partner is expected to send at least one representative to each review and scientific meeting. The locations for these meetings will normally be chosen in rotation from among the 6 member sites, or elsewhere if the Commission requests it. More limited technical meetings and exchanges between the partners will be arranged as required, to ensure a strong scientific collaboration.

Review meetings will be held either at one of the partner sites, or at a central location, as agreed with the Commission. For all meetings, the Coordinator will send an agenda well before the meeting.

All partners must promptly deliver any financial, legal or technical information relating to the project, that is requested by the Coordinator. The Coordinator will make sure that deadlines for submission of the information will be reasonable, excluding *force majeure*. The Coordinator will notify the availability of deliverables to all Partners as soon as they reach the Coordinator and a list of available public deliverables will be updated and made available through WWW to third parties.

9.7.2 Assessment and evaluation

Work package 9 is devoted to assessment and evaluation. Each scientific WP leader is responsible for monitoring the progress and conformance to the workplan of his/her workpackage, for reporting the results to the project coordinator, and if necessary for recommending suitable remedial actions. The evaluation criteria are as stated in each WP description. Assessment reports must be made as soon as any substantial deviation from the package workplan is noticed, and at a minimum every 6 months. The coordinator will: review the overall progress of the project at least every 6 months, or whenever serious problems or delays arise; ensure that appropriate remedial action or fall-back plans are discussed by the project management committee; and monitor the implementation of any adopted workplan changes. The coordinator will also issue annual executive summaries of the current project status (deliverables D 9.1–9.3), and inform the Project Officer promptly of any major delays or changes in workplan.

As an example of the sorts of actions that might be taken, WP 6.2 requires an initialization in order to track walking humans. If extracting spatial regions of walking humans using motion segmentation proves infeasible in WP 1.2, then more person months will be devoted to recognizing humans using prototype shapes and statistical distributions in WP 2.2.

10. Clustering

VIBES is not part of an identified cluster of current research projects, so no explicit clustering plans have been made.

11. Other Contractual Conditions

11.1 Subcontractors

There are no subcontractors on the project.

11.2 Travel Outside EU Member and Associated States

The following VIBES partners wish to devote the following fractions of their travel budgets towards presenting their VIBES-related results at major international conferences held outside the European Union and Associated States:

Partner	**Year 1** Euro	**Year 2** Euro	**Year 3** Euro	**Total** Euro	**% Total Site Travel Budget**
KTH	6000	5500	6000	17500	45%
Weizmann	7000	7000	7000	21000	47%
INRIA	3000	3000	3000	9000	30%
K.U.Leuven	2000	2000	2000	6000	36%
Oxford	1300	1300	1300	3900	20%

These presentations will ensure a high international profile for VIBES research, and may also encourage contacts with implementors and end-users wishing to license VIBES technology. Moreover, attendance at such conferences is essential to keep abreast of the rapid international development of video technologies relevant to VIBES.

The non-European conferences that the partners are most likely to attend are the International Conference on Computer Vision (ICCV), which will be held in Toronto in 2001 and in Beijing in 2003, and Computer Vision and Pattern Recognition (CVPR), which is held annually in the United States or Canada. Along with the major conference inside Europe — the bi-annual European Conference on Computer Vision, which all members of the VIBES consortium support and attend very strongly — these are the most prestigious conferences in the computer vision domain, and both of them are highly relevant to VIBES' themes.

The cost of travel to such conferences is highly dependent on the location (major cities being more expensive owing to increased hotel and conference facility costs) and the availability of discounted airfares. For a major conference in a large city outside Europe: flights are likely to cost at least 1000–1500 Eu (somewhat more for travel from Israel); registration for the main conference plus one workshop should cost about 600–700 Eu for a full time researcher, and about 350 Eu for a student; and a safe, clean reasonably central hotel is likely to cost between 300 and 600 Eu (6 nights 50–100 Eu). So the total cost for one person is in the range of 2000–3000 Eu. Hence, KTH and Weizmann are asking support for about 2–3 international person conferences per year, INRIA for about 1–1.5 and Leuven and Oxford for about 0.5. (INRIA, Leuven and Oxford will all supplement this requested support with their own internal travel funds).

11.3 Protection of Knowledge and Other Specific Costs

The consortium will adopt the provisions for protection of knowledge in the model contract.

No other specific costs are foreseen.

A. Consortium Description

A.1 The Consortium

The consortium includes six research groups: KTH, Sweden; the Weizmann Institute of Science, Israel; INRIA Rhône-Alpes, France; K.U. Leuven, Belgium; University of Oxford, UK; and EPF Lausanne, Switzerland. All of these are experienced, internationally known computer vision groups, who have managed or contributed to many successful collaborative projects in the past. They have been selected for their complementarity of interests and experience, and for the relevance of these skills to VIBES goals.

KTH will focus on recognition of dynamic object classes in scenes, especially human motion and activity. The human dynamic information in the image will first be localized, then the projected shape will be analyzed over time in order to assess the nature of the motion and classify it w.r.t. pre-specified or learned activity categories.

Weizmann will generalize their work on mosaics towards more general motions and scenes. They will also distinguish between the static and dynamic parts of the scene. The work will include the determination of precise motions over multiple frames, *e.g.* to distinguish between rigid and articulated motion. Finally, they will semantically analyze video clips according to identified object categories.

INRIA will bring in expertise on video browsing tools, and generalize their methods of quasi-invariant feature based object recognition for the detection of different shots of the same scene. They will also contribute experience in bundle adjustment technology to the 3D reconstruction and model-building work.

Leuven will build 3D dense reconstructions of scenes. Depth information and regularities in the scene will be used to perform scene type classification. Leuven will also detect shots of the same scene, based on the matching of geometric/photometric scene patches. They will extend the classes of such patches. Leuven will also detect human faces as a shape class of special interest.

Oxford will compute cameras and scene reconstructions for sequences. As a second topic, they will bring their wide baseline work to bear on the detection of similar scenes in different shots. Oxford will also generalize previous work on grouping which will be used in the classification of scene types.

EPFL will track humans, based on detailed kinematic models of the human body. Based on the extracted body motions, they will classify human actions. This model based approach is complementary to the non-kinematic method employed by KTH. EPFL will also compare people in different scenes, in order to find the same person elsewhere. They bring experience in contributing to the MPEG-4 standard.

A.2 Description of the Participants

Partner 1: KTH - Computational Vision and Active Perception Laboratory

The CVAP (Computational Vision and Perception Laboratory) performs research in computer vision and robotics since 1982. The group was formed in 1982 and has today 30 researchers. It is associated with the department of numerical analysis and computing science at KTH, Stockholm.

The research is currently mainly funded through two Swedish sources, TFR, the Swedish Research for Engineering Science, and SSF, the Swedish Foundation for Strategic Research and through a set of grants from the European Union. CVAP has since 1993 a long-term so-called frame grant from TFR and is a partner in two consortia sponsored by SSF, CAS: The Center for Autonomous Systems, and VISIT: Visual Information Technology. VISIT is a national research program hosted by Uppsala University, and involves groups from six Swedish universities. Jointly with the Karolinska Institute, CVAP has a project on the analysis of functional brain images.

Up to 1995 CVAP were partners in three ESPRIT Basic Research Actions: Insight II, VAP II and VIVA. After the ending of these Frame Program III projects, CVAP currently participates in the ESPRIT LTR project Improofs and three networks: RETINA, VIRGO and CAMERA. The group is also a primary node in ECVNet, the European Computer Vision Network.

Stefan Carlsson

Stefan Carlsson joined the computer vision group at KTH in 1991. Between 1992 - 1995 he participated in the European Esprit Basic Research Action 6448, VIVA which was a consortium of some of the major computer vision groups in Europe for doing research on invariance in computer vision. In 1997 he started work in collaboration with partners from the VIVA project in the program IMPROOFS, which involves application of computer vision to to problems in forensic science. He is also involved in the national program VISIT, as a national coordinator for the project “view-synthesis” and also in the project “content based search in image databases”.

References

1. S. Carlsson, Duality of reconstruction and positioning from projective views, In *Proceedings of the IEEE Workshop on Representations of Visual Scenes*, Cambridge, Mass, June (1995).

2. S. Carlsson, Projectively Invariant Decomposition and Recognition of Planar Shapes, *International Journal of Computer Vision*, Vol 17, No 2, pp. 193 - 209 Feb. (1996a).

3. S. Carlsson, Combinatorial Geometry for Shape Representation and Indexing, *Object Representation in Computer Vision II*, Springer Lecture Notes in Computer Science 1144, pp. 53 - 78 Ponce, Zisserman eds. (1996c).

4. S. Carlsson, Geometric Structure and View Invariant Recognition, *Philosophical Transactions of the Royal Society of London*, Series A 356, 1233 - 1247 (1998).

5. S. Carlsson, Order Structure, Correspondence and Shape Based Categories, *International Workshop on Shape Contour and Grouping*, Torre Artale, Sicily, May 26 - 30 1998, Springer LNCS (to appear).

Partner 2: Weizmann Institute of Science

The vision group within the department of Computer Science and Applied Mathematics at the Weizmann Institute of Science is a research team of 15-20 people, consisting of PhD and MSc students, and 3 faculty members: Prof. Shimon Ullman, Dr. Ronen Basri, and Dr. Michal Irani. Its research covers various problems in computer vision, including segmentation and perceptual grouping, video and motion analysis, video applications, recovery of 3D geometry, object recognition, image database retrieval, and robot navigation. The team participates in projects funded by the Israeli Ministry of Science, Israeli Science Foundation, the Israeli Ministry of Defense, and the Minerva Foundation, and has been funded in previous years by DARPA, Israel-US Binational Science Foundation, and the Israeli Ministry of Trade and Commerce.

Michal Irani

Michal Irani joined the Department of Computer Science and Applied Mathematics in 1997, where she has been appointed as a Senior Scientist (Senior Lecturer). In 1994 Dr. Irani received the David Sarnoff Research Center Technical Achievement Award for her work on ”the development of mosaic representations and techniques for the visualization, compression, indexing, and editing of video”. In 1998 she received the Yigal Allon 3-year fellowship awarded to outstanding young scientists by the Israeli Council for Higher Education. Dr. Irani is an associate editor in the major international Computer Vision journal (IEEE T-PAMI), and a member of the program committees of several international conferences in Computer Vision and Computer Graphics (CVPR and SIGGRAPH). Dr. Irani is the Principal Investigator on grants from the Israeli Ministry of Science, the Israeli Science Foundation, the Israeli Ministry of Defense, and the American agency DARPA.

Ronen Basri

Ronen Basri joined the Department of Computer Science and Applied Mathematics in 1992 where he has been appointed as a Senior Scientist. His research focuses on high level visual tasks such as object recognition and categorization and their application to image database retrieval and robot navigation. In this research Ronen Basri has attempted to show how direct information of visual knowledge that is extractable from images can be used to solve complex visual tasks. His expected to contribution to the VIBES project is in developing methods for semantic analysis of video data. Ronen Basri is an associate editor in one international computer vision journal and a member of the program committees of several international conferences in the field. His research has been funded by the Israeli Ministry of Science, the Israeli Ministry of Trade and Commerce, the Israeli Science Foundation, the Israel-US Binational Science Foundation, and the Minerva Foundation.

References

1. M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu, Efficient representations of video sequences and their applications, Signal Processing: Image Communication, special issue on Image and Video Semantics: Processing, Analysis, and Application, **8**(4), 1996.

2. M. Irani and P. Anandan, Video indexing based on mosaic representations, Proc. of IEEE:905–921, 1998.

3. M. Irani and P. Anandan, A unified approach to moving object detection in 2D and 3D scenes, IEEE Trans. on Pattern Analysis and Machine Intelligence, **20**(6):577–590, 1998.

4. M. Irani, P. Anandan, and D. Weinshall, From Reference Frames to Reference Planes: Multi-View Parallax Geometry and Applications, European Conference on Computer Vision, 1998.

5. M. Irani, Multi-Frame Estimation of Optical Flow Using Subspace Constraints, IEEE International Conference on Computer Vision, 1999, forthcoming.

6. Ronen Basri, Recognition by prototypes, *International Journal of Computer Vision*, **19**(2):147–168, 1996.

7. Ronen Basri, Luiz Costa, Davi Geiger, and David W. Jacobs, Determining the similarity of deformable shapes. *Vision Research*, **38**:2365–2385, 1998.

8. R. Basri, D. Roth, and D.W. Jacobs, Clustering appearances of 3D objects, *IEEE Conf. on Computer Vision and Pattern Recognition*:414–420, 1998.

9. R. Basri and Y. Moses, When is it possible to identify 3D objects from single images using class constraints, *International Journal of Computer Vision*, **33**(2), 1999.

10. Ronen Basri, Ehud Rivlin, and Ilan Shimshoni, Visual homing: surfing on the epipoles. *International Journal of Computer Vision*, **33**(2), 1999.

Partner 3: MOVI, INRIA Rhône-Alpes

MOVI (MOdelling for VIsion, `http://www.inrialpes.fr/movi`) is a vision research team of about 22 people including 5 permanent researchers, lead by Dr Radu Horaud. It is located at INRIA Grenoble, and is part of the GRAVIR (Graphics, Vision and Robotics) grouping which combines researchers from INRIA, CNRS, the Institut National Polytechnique de Grenoble, and the Université Joseph Fourier.

MOVI's research focuses on geometric aspects of vision, especially 3D reconstruction, geometric invariants, recognition and image databases, and vision based robotics. Its work spans both theoretical and applied aspects.

MOVI currently leads the Esprit LTR projects 21914 CUMULI and 26247 VIGOR, and has successfully participated in many other national and international joint research projects in the past, including the Esprit projects VIVA, FIRST, SECOND, ARVISA, VIMINI, the TMR programme and the DARPA-ESPRIT action in vision for geometry. It also participates in the French GDR's (industrial-academic research groupings) "Man-Machine Communication" (ORASIS project) and "Analysis of Signals and Images" and has held contracts or cooperations with the companies ITMI, MATRA and AEROSPATIALE. It has had collaborations and exchanges with many other local and international research groups, recently Illinois, Xi'an, Ljubljana, Gunma, Oxford, Melbourne.

Cordelia Schmid

Cordelia SCHMID is an INRIA permanent researcher (Chargée de Recherche) in the MOVI team at INRIA Rhône-Alpes. Her work centres on local geometric invariants for feature correspondence and indexing into image and video databases. She holds a doctorate from the Institut National Polytechnique de Grenoble (MOVI team). She recently held a post-doctoral position at Oxford University, working on 3D reconstruction of buildings from sequences of aerial images for Esprit LTR project 20243 IMPACT.

Bill Triggs

Bill TRIGGS is a contract researcher at INRIA, working on and managing the Esprit LTR photogrammetry and computer vision project 21914 CUMULI (Computational Understanding of Multiple Images). He is a mathematical physicist turned robotician turned vision researcher. Since 1995 he has worked mainly on geometric aspects of 3D vision including visual reconstruction and invariants. This includes both theoretical foundations (geometry and tensorial formulation of multi-image matching relations, theory of self-calibration) and numerical and statistical aspects (efficient structure estimation, noise modelling).

References

1. C. Schmid and R. Mohr, Local Grayvalue Invariants for Image Retrieval, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **19**(5):530–534, 1997.

2. C. Biernacki and R. Mohr, *Indexation et Appariement d'Images par Modèles de Mélanges Gaussien des Couleurs*, INRIA RR-3600, 1999.

3. R. Mohr, P. Gros and C. Schmid, Efficient Matching with Invariant Local Descriptors, *Proc. of the Joint IAPR International Workshops SSPR98 and SPR98: Advances in Pattern Recognition*, **1451**, Sydney, Australia, Lecture Notes in Computer Science, Springer-Verlag:54–71, 1998.

4. P. Bouthemy, Y. Dufournaud, R. Fablet, R. Mohr, S. Peleg and A. Zomet, Video Hyperlink Creation for Content-Based Browsing and Navigation, *Workshop on Content-Based Multimedia Indexing, CBMI'99*, Toulouse, France, 1999.

5. P. Sturm and B. Triggs, A Factorization Based Algorithm for Multi-Image Projective Structure and Motion, *European Conf. on Computer Vision*, Cambridge, U.K.:709–720,1996.

6. B. Triggs, Optimal Estimation of Matching Constraints, *Workshop on 3D Structure from Multiple Images of Large-scale Environments SMILE'98*, R. Koch and L. Van Gool (Eds.), Lecture Notes in Computer Science, Springer-Verlag, 1998.

7. B. Triggs, Differential Matching Constraints, *IEEE International Conference on Computer Vision*, 1999, forthcoming.

Partner 4: K.U. Leuven

The team working on the project at the University of Leuven is the group VISICS (VISion for Industry, Communications, and Services) which is part of the Center for the Processing of Speech and Images of the university's department of Electrical Engineering (ESAT). With its 25 researchers — 7 of whom are postdoc — it covers several areas of computer vision. These include remote sensing, visual inspection, shape reconstruction and recognition, robot vision, and image compression. Over recent years the group has received several prizes for its research, including a David Marr Prize, a Henry Ford European Conservation Award, two TechArt prizes, Barco best dissertation awards, etc.

Especially relevant for the project is work on (1) 3D reconstruction from uncalibrated video sequences and (2) matching under wide baseline conditions using geometric/photometric invariant descriptions of scene patches. Both of these strands will be used in the project, which will call for substantial extensions over the current state of the art in each of these lines of research.

Luc Van Gool

Prof. Luc Van Gool, born in 1959 in Antwerp Belgium, will lead the research carried out under this project. He is professor at the University of Leuven since 1992 (1992 Assistant Prof., 1994 Associate professor, and 1996 professor). He leads the computer vision group VISICS. The creation of 3D models with simple means and the use of invariant description of both 2D and 3D shapes and patterns are among his major areas of interest. In 1998 he received the David Marr Prize, together with his colleagues Marc Pollefeys and Reinhard Koch. Also in 1998 he was the co-founder of the spin-off company Eyetronics. Luc Van Gool has worked in several European projects before and has coordinated 4 European projects (3 Esprit and 1 Acts). He is a member of the programme committees of several international conferences in the field of computer vision (a.o. the European Conference on Computer Vision and the International Conference on Computer Vision).

References

1. M. Pollefeys, R. Koch, and L. Van Gool, Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters, Int. Conf. on Computer Vision, pp. 90-95, Bombay, India, jan. 4-7, 1998.

2. R. Koch, M. Pollefeys, and L. Van Gool, Automatic 3D model acquisition from uncalibrated image sequences, Proc. Computer Graphics International, pp. 597-604, Hannover, 1998.

3. T. Tuytelaars, L. Van Gool, L. D'haene, R. Koch, Matching Affinely Invariant Regions for Visual Servoing, International Conference on Robotics and Automation, Detroit, pp. 1601-1606, May 10-15, 1999.

4. T. Tuytelaars, L. Van Gool, Content-based Image Retrieval based on Local Affinely Invariant Regions, submitted to the International Conference on Visual Information Systems, Visual99, Amsterdam, June 2-4, 1999.

5. L. Van Gool, R. Koch, and T. Moons, New techniques for 3D modeling... and for doing without, Proc. 6th Int. Symp. on Experimental Robotics, pp. 55-64, Sydney, 1999.

Partner 5: Robotics Research Group, University of Oxford

The VIBES project will be carried out by the Visual Geometry Group within the Robotics Research Group (RRG) of the Department of Engineering Science. The RRG is one of the largest and best known in its field in Europe, with five faculty and around fifty researchers in total. The group has been involved in a number of previous Esprit BRA's (FIRST, INSIGHT, SECOND, INSIGHT-II, VIVA,IMPACT) as well as ACTS Project AC074 VANGUARD. It is currently involved in the Esprit LTR IMPROOFS. Its faculty have consulted widely for major companies such as Siemens, IBM, GE and Sharp, and enjoy extensive industrial support for their research. Government support has included substantial grants from the Engineering and Physical Sciences Research Council, the Department of Trade and Industry and the UK Defence Research Agency.

The RRG has extensive experience with standard computer vision and image processing techniques through applications ranging from satellite images (determining rural areas) through to inspection of agricultural products (detecting weeds). Most relevant to VIBES are (1) uncalibrated reconstruction – where 3D models are built automatically from uncalibrated image sequences; (2) wide baseline matching – where images from very different viewpoints of the same scene are automatically matched.

Andrew Zisserman

Prof. Andrew Zisserman joined the Department of Engineering Science, University of Oxford in 1987, and is a University Research Lecturer. He leads the Visual Geometry Group. He is Principal Investigator at Oxford for EC Esprit Project Improofs, and for three further research grants funded by UK research agencies. He has authored over 90 papers in Photogrammetry and Computer Vision. He is on the editorial board of two international Computer Vision journals, and on the IEEE PAMI awards committee. He has recently been program co-chair for the International Conference on Computer Vision, and an area chair for the Conference on Computer Vision and Pattern Recognition. He was awarded the IEEE Marr Prize in 1998, together with Dr. Fitzgibbon and Dr. Torr at Oxford, for their work on estimating multiple view geometry. This work is directly relevant to the VIBES project.

References

1. P. A. Beardsley, P. H. S. Torr and A. Zisserman, 3D Model Acquisition from Extended Image Sequences, *European Conf. on Computer Vision*:683–695, 1996.

2. C. Baillard and A. Zisserman, Automatic reconstruction of piecewise planar models from multiple views, *IEEE Conf. on Computer Vision and Pattern Recognition*:559–565, 1999.

3. P. Pritchett and A. Zisserman, Matching and reconstruction from widely separated views, In R. Koch and L. Van Gool, editors, *3D Structure from Multiple Images of Large-Scale Environments*, Lecture Notes in Computer Science 1506, Springer-Verlag:78–92, 1998.

4. C. Schmid, A. Zisserman, and R. Mohr, Integrating geometric and photometric information for image retrieval. In D.A. Forsyth, V. Di Gesu, J.L. Mundy, and R. Cipolla, editors, *Shape, Contour and Grouping in Computer Vision*, Lecture Notes in Computer Science, Springer-Verlag, 1999.

5. P. H. S. Torr, A. W. Fitzgibbon and A. Zisserman, Maintaining Multiple Motion Model Hypotheses Over Many Views to Recover Matching and Structure, *IEEE International Conference on Computer Vision*:485–491, 1998.

Partner 6: Ecole Polytechnique Fédérale de Lausanne (EPFL)

The Computer Graphics Lab (LIG) at the Swiss Federal Institute of Technology in Lausanne was created in July 1988 and completely organized by its director, Professor Daniel Thalmann. The laboratory includes a staff of 32 people.

Research at LIG is oriented towards the simulation of Autonomous Virtual Humans in Virtual Worlds. We emphasize the aspects of motion control and the simulation of high-level behaviours based on physical laws, artificial intelligence, and autonomous agent technology. We develop models for autonomy and perception based on real and virtual sensors: visual, auditive, and tactile. We also develop multimodal interactions with Virtual Humans.

There are 6 main Research areas at LIG: Motion Control, Artificial Life and Behavioral Animation, Shared Virtual Environments, Computer Vision and Augmented Reality, Body deformations, and Standards for body description and animation (VRML, MPEG-4).

In particular, we extensively work on gesture and action recognition in order to direct virtual humans. In order to recognize motion, we use magnetic motion capture and video-based tracking. We are currently extending our models handle to groups of Virtual Humans, crowds, and interaction with smart objects such as elevators or escalators. A multimodal interface allows us to guide crowds and we are developing a rule-based system autonomous crowds.

EPFL developed the MPEG-4 reference software for body animation, and donated this software and corresponding models to ISO for distribution. The work has been coordinated with MIRALab University of Geneva who donated the face animation software. LIG-EPFL also hosts the MPEG-4 body animation test data set.

In cooperation with MIRALab (University of Geneva), we also work on shared Virtual Environments, where avatars and autonomous virtual humans can meet in common virtual spaces. There are applications of this technology in entertainment, teleconferencing, and medicine.

LIG has a strong international cooperation through several European projects (Esprit, ACTS, Biomed...) in the areas of Computer Vision, Virtual Reality, Shared Environments, Computer Animation, and Medical applications and is also involved in 4 national projects. It is currently involved in a number of IST projects. In particular, along with K.U. Leuven, it is a member of the *MESH — Modeling of Expression and Shape of human Heads* consortium.

The laboratory is sponsored by the Swiss Federal Institute of Technology, the National Swiss Research Foundation, and the Federal Office for Education and Science. It is well equipped with PCs and SGI workstations including an Onyx2 Infinite Reality 2. We have also digital facilities for video and audio and various VR devices.

Pascal Fua

Pascal Fua joined LIG in 1996 as an Assistant Professor (Maitre d'Etudes et de Recherches). Before that, he worked at SRI International and at INRIA Sophia-Antipolis as a computer scientist. He holds a degree from Ecole Polytechnique, Paris, and a Ph.D. in Computer Science from the University of Orsay. He has (co)authored over 50 publications in refereed journals and conferences. He is a member of the programme committees of the European Conference on Computer Vision.

Daniel Thalmann

Daniel Thalmann is a full Professor and the Director of the Computer Graphics Laboratory at EPFL. He is coeditor-in-chief of the Journal of Visualization and Computer Animation and member of the editorial board of several journals. Daniel Thalmann was member of numerous Program Committees, Program Chair and Conference Chair of several conferences. He will serve as Program Co-chair for IEEE VR 2000. He has also organized 4 courses at SIGGRAPH on human animation. Daniel Thalmann is a pioneer in research on Virtual Humans. He has published more than 200 papers, is coeditor of 25 books, and co-author of several books on Computer Animation, Image Synthesis and Networked Virtual Environments.

Ronan Boulic

Ronan Boulic is a Senior Researcher, Lecturer and PhD Director at the Computer Graphics Lab of the Swiss Federal Institute of Technology, Lausanne (EPFL). He received the PhD degree in Computer Science in 1986 from the University of Rennes and the Habilitation degree from the University of Grenoble, in 1995. Ronan Boulic is co-author of 45 research papers in international conferences and journals. He was chair of the Eurographics Workshop on Computer Animation and Simulation 1996 and co-editor of the associated book published by Springer Verlag. He is member of the program committee of Eurographics'CAS (since 93), Computer Animation'99 and CGIM'99.

1. P. Fua. Regularized Bundle-Adjustment to Model Heads from Image Sequences without Calibration Data. *International Journal of Computer Vision*. In Press.

2. P. Fua, L. Herda, R. Plänkers and R. Boulic, Human Shape and Motion Recovery Using Animation Models. In *XIX ISPRS Congress*, Amsterdam, Netherlands, July 2000.

3. P. Fua. Using Model-Driven Bundle-Adjustment to Model Heads from Raw Video Sequences. In *International Conference on Computer Vision*, Corfu, Greece, September 1999.

4. P. Fua, R. Plänkers, and D. Thalmann. From Synthesis to Analysis: Fitting Human Animation Models to Image Data. In *Computer Graphics International*, Canmore, Alberta, Canada, June 1999.

5. R. Plänkers, P. Fua, and N. D'Apuzzo. Automated Body Modeling from Video Sequences. In *ICCV Workshop on Modeling People*, Corfu, Greece, September 1999.

6. T. Capin, I. Pandzic, N. Magnenat-Thalmann, and D. Thalmann, *Avatars in Networked Virtual Environments*, John Wiley and Sons, 1999.

B. Contract Preparation Forms

The CPF’s go here.