



Sixmonthly report

No: 7 (period 9)

From *February 1, 2008 to July 31, 2008*

Project Start:	01-02-2007
Project Duration:	36 Months
Priority area	2.3.6
Contract No.:	FP6-045547
Website:	www.vidivideo.info

Due-Date:	15-09-2008
Delivery:	15-09-2008
Lead Partner:	University of Amsterdam
Project Leader	Prof. Arnold Smeulders
Dissemination Level:	Restricted-PP
Status:	Final
Approved:	
Version:	1.0

6-monthly Management Report Vidi Video

9th Period

Project months: Months 13 to 18
Febuari 1 – July 31, 2008

Project Ref. no.:	045547
Acronym:	Vidi Video
Reporting period:	February 1 to July 31, 2008
Project coordinator*:	Prof. Arnold Smeulders
Tel:	+31-20 525 7460
Fax:	+31-20 525 7490
Email:	a.w.m.smeulders@uva.nl
Website:	www.vidivideo.eu

1. Progress

Work done:

WP1: Project Management

Workpackage leader: UVA

In the sixmonth period from February 1 to July 31, 2008 the following things were organised.

The third VidiVideo meeting was held on February 7 in Barcelona. At this meeting the annual report and review were planned. It was a very fruitful meeting. At this meeting a governing board meeting was held too.

The first Periodic Report consisting of a Management report and an Activity report were delivered.

On March 18 a preparatory meeting for the VidiVideo Review was held.

On March 19 the First year Review of the project was held. Stefano Bertolo (scientific officer of the EU) as well as Enis Cetin (reviewer, Belkint University) were present.

Year 1 Review of project VidiVideo was formally declared successful by the officer of the EU. There were a few details that required attention. We made adjustments to the financial part of the periodic Management report. And we improved two deliverables on the basis of the outcome of the review.

A teleconference was held om May 23, 2008 to get the second year of the project going. Of all the partners at least one person was attending.

First mentioned partner is the work package leader



From March on Caroline van Impelen has joined the VidiVideo management team.

In June the payment of the Community financial contribution was received. And the payment of the Community financial contribution was distributed to the partners/contractors.

The next VidiVideo meeting will be held in Lisbon - Portugal on October 9 and 10 2008.

WP2: Video Processing: Workpackage leader **CERTH**
Co-reporting partners: Inesc-ID, UvA

- *Progress towards objectives:*

1: *Scientific & technical aspects;*

During this reporting period WP2 partners continued working on efficient video shot and audio segmentation and representation, in accordance with the project objectives. Recent developments in shot segmentation include starting and completing the development of a new gradual transition detection approach with very good results, which have already lead to the authoring of two relevant conference papers that have been accepted. Concerning shot representation, work has began on shot grouping, on which the relevant literature has been studied and relevant implementation work is about to begin. Joint audio-video processing approaches are examined towards this. Recent developments in audio segmentation include streamlining the procedure for training models for new anchors and modifying the output format of the audio segmentation module to make it compatible with the MPEG-7 format adopted in the project. Work also started on a 3-class gender detector, including children as well as adult male and female voices, and on improving our music detection method. D 2.2 Video processing software, first version was also completed and delivered at the beginning of this 6-month reporting period.

- *Work Performed:*

Task 2.1 Video Segmentation

CERTH, based on the results of the first year, we focused in this period on the development of a new gradual transition detection approach. The new approach is based on the observation that the algorithms in the literature for gradual transition detection have moderate performance, mostly due to their sensitivity to global or local motion, illumination changes etc. present in real video sequences. In the new approach under development, two main directions are followed to help overcome these limitations. As a first direction, novel shot change detection features are proposed, exhibiting reduced sensitivity to motion activity. In this period, three such features were developed: one based on the Macbeth color pallet difference, one based on color coherence change, and a third one based on the luminance center of gravity difference between adjacent frames. As a second direction, a method of combining them in a meta-segmentation scheme, in order to achieve more accurate detection results, is proposed. In its latest update, this involves considering the aforementioned shot change detection features not only between pairs of temporally consecutive frames, but extending them in a temporal multi-resolution setting (i.e. also considering the differences between non-adjacent frames that are 2, 4, 8 etc. frames apart) that allows more accurate detection of gradual transitions. A thorough evaluation and comparison of this novel scheme was conducted and showed superior results over several methods of the relevant literature that were applied to our test corpus. **CERTH** also lead the authoring of D 2.2 “Video processing software, first version”, which was completed and delivered at the beginning of this 6-month reporting period.

Inesc_ID: Not involved in this task.

UvA: Not involved in this task.

Task 2.2 Shot representation, key frames

CERTH started work on shot grouping, as part of the overall shot representation approach in Vidi-Video. To this end, CERTH studied in this period literature approaches for the meaningful grouping of shots, including multimodal ones. A section on joint audio-video processing (thought not specific on shot grouping; instead, examining the more generic issue of the usefulness of joint audio-video processing in different video analysis and representation tasks) was also added to D2.1, in response to a relevant reviewer comment. After a thorough review of the literature, relevant implementation work is about to begin. This will result in a shot grouping method for Vidi-Video within the following reporting periods.

Inesc_ID: Not involved in this task.

UvA: Nothing to report in this period

Task 2.3 Audio segmentation

Inesc_ID: During the first 2 months of the 6-month period, the work of INESC-ID in terms of audio segmentation was devoted on one hand to streamlining the procedure for training models for news anchors, and on the other hand to modifying the output format of our audio segmentation module to make it compatible with the MPEG-7 format for audio and video segmentation adopted in the project. A demo of the joint work of WP2 partners (shot + audio segmentation) and Task 3.2 (speech recognition) was also prepared for the review meeting.

Subsequently, the speech-non-speech detector was enhanced with a new world model. This new model now includes a wide representation of very different audio events, extracted from the sound effect corpus collected in WP3. An adaptation process, after the normal training of the classifier, allowed a compensation of the amount of training in the different classes. This contributed for instance to a better rejection of speech with background jingle music, which seriously degraded the speech recognizer.

The team also invested on building an AGC module (Automatic Gain Control). Both the audio segmentation and the speech recognition modules had a worse performance when fed with low amplitude audio signals. Besides amplitude normalization, this module also compensates the amplitude of the signal in cases where a low-frequency modulation distorts the input signal.

Inesc_ID also started work on a 3-class gender detector: male, female and child. This implied using a corpus of children voices, available from LDC (The CMU Kids Corpus). Unfortunately, the size of this corpus is significantly smaller than the one originally used for training the male/female detector. This implied a severe reduction of the size of the male/female training corpus, in order to adequately train an MLP with PLP features extracted from this more balanced corpus. Another limitation of the children voices corpus is that it only includes voices from children from the first, second and third grades. Despite these limitations, the detector achieved 89.9% correction on the cross-validation corpus. No adult voices were classified as children, but some children voices were classified as adult, and the discrimination between adult voices naturally degraded due to the much lower amount of data available for training. On a subset of the children voices corpus reserved for testing, the detector achieved an error of 2.85%.

We also started working on improving our music detection method. This started by a literature survey and involved the creation of a corpus of 162 files with non-vocal music, totalling almost 15 h, and a corpus of 134 files with vocal music, totalling around 9 h. Preliminary experiments with autocorrelation based features did not yield yet good results. Although not exactly part of the work on this project, our complementary work on topic identification and spoken language identification was also useful to the goals of this audio segmentation task.

CERTH: Not involved in this task.

UvA: Not involved in this task.



- *Deviations*

None

- *Dissemination activities:*

CERTH:

Attended the VidiVideo meeting in Barcelona (Feb. 2008) and the 1st VidiVideo Review in Amsterdam (March 2008).

Relevant publications:

E. Tsamoura, V. Mezaris, I. Kompatsiaris, "Video shot meta-segmentation based on multiple criteria for gradual transition detection", Proc. Sixth International Workshop on Content-Based Multimedia Indexing (CBMI 2008), London, UK, June 2008.

E. Tsamoura, V. Mezaris, I. Kompatsiaris, "Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework", IEEE International Conference on Image Processing, Workshop on Multimedia Information Retrieval (ICIP-MIR 2008), San Diego, CA, USA, to appear.

INESC-ID:

Attended the VidiVideo meeting in Barcelona (Feb. 2008) and the 1st VidiVideo Review in Amsterdam (March 2008).

Relevant publications:

Hugo Meinedo, Márcio Viveiros, João Neto, Evaluation of a Live Broadcast News Subtitling System for Portuguese, INTERSPEECH 2008, Brisbane, Australia, September 2008.

Rui Amaral, Isabel Trancoso, Topic Segmentation in a Media Watch System, PROPOR 2008, Aveiro, Portugal, September 2008

Rui Amaral, Isabel Trancoso, Topic Segmentation and Indexation in a Media Watch System, INTERSPEECH 2008, Brisbane, Australia, September 2008

Jean-Luc Rouas, Isabel Trancoso, Céu Viana, Mónica Abreu, Language and Variety Verification on Broadcast News for Portuguese, Speech Communication, online publication complete: 10-JUL-2008, DOI information: 10.1016/j.specom.2008.05.006

UvA:

Attended the VidiVideo meeting in Barcelona (Feb. 2008) and the 1st VidiVideo Review in Amsterdam (March 2008).

Status of Deliverables, by July 31, 2008

Deliverable	WP Leader	Status: completed / under way / not started yet		On schedule yes/no	Original completion date		Actual completion date calendar date
start date						1-feb-2007	
D2.1	CERTH	completed		yes	month 12	1-feb-2008	1-feb-2008
D2.4	CERTH	completed		yes	month 12	1-feb-2008	1-feb-2008
D2.5	CERTH	completed		yes	month 12	1-feb-2008	1-feb-2008
D2.3	CERTH	under way		yes	month 28	1-jun-2009	
D2.2	CERTH	completed		yes	month 14	1-apr-2008	1-apr-2008

WP3: Audio Analysis: Workpackage leader **INESC-ID**
Co-reporting partners: CERTH

- *Progress towards objectives:*

1: *Scientific & technical aspects;*

During this reporting period WP3 partners continued working on audio event detection and on speech recognition, in accordance with the project objectives. Research efforts in term of audio event detection were aimed at training one-against-all SVM classifiers for semantic concepts with high representativeness in the training corpus of sound effects. At an early stage, due to the very small pilot corpus, the list of concepts was very reduced, but already enabled the first integration of the detection results in the application designed by the University of Amsterdam, that allows queries to a video database. At a later stage, the availability of an extended training corpus enabled the training of new concept detectors, and their exhaustive testing with many different combinations of features, kernels, etc. In terms of speech recognition, the research efforts were first aimed at evaluating the current performance with publicly available corpora for US English. Subsequently, the work focused on a two-pass approach that takes into account the confidence measure of each word recognized during the first pass, and adapts the acoustic models of each speaker based on the words recognized with confidence higher than a threshold.

- *Work Performed:*

Task 3.1 Detection of Audio Events

INESC-ID: During the first 2 months of this period, the work of INESC-ID in terms of audio segmentation, covered several directions: we trained new one-against-all SVM classifiers for three concepts (machines working, water, walking/climbing); we tested different feature combinations, including features that explore the high frequency contents of the signals; we tested several window sizes and update frequencies; we experimented several thresholding techniques for discarding very low energy frames; and we addressed the problem of the speed of SVM training, by using scaling and restricting probability estimates to final classifiers. Of particular importance was the work towards the integration of the MPEG-7 files produced by this module in the application designed by the University of Amsterdam, that allows queries to a video database by means of a GUI, created using OpenGL libraries. This integration work also involved the development of a post-processing stage. In order to test the application, we manually labelled 3 documentary files provided by FRD in terms of the presence or absence of one of the trained concepts (birds). Results were very promising.

We also have actively cooperated with B&G in defining the full sound effect corpus required for the extensive training of new concept detectors. This corpus was received in early May, comprising 290h of audio. Running the audio segmentation module on a total of 920 files, 10.4% were classified as containing some speech. The percentage of time of detected speech was 1.4%. That motivated retraining the non-speech model with a mixture of these sound effects, besides the jingles and other data previously included, such as jingles. With the new world model, only 0.8% of the files were still classified as containing speech. The percentage of time of detected speech was 0.07%.

With the extended corpus, experiments were run on a wide range of semantic concepts, in order to select the best kernel (linear, polynomial, radial basis function), and the best feature combination. The set of concepts was: Airplane (jet and propeller), Bird, Bus, Cat-Meowing, Crowd-Applause, Dog (barking), Gunshot, Helicopter, Horse (walking), Telephone-Ringing (bell and digital), Traffic, Sirens, and Water. On the cross-validation set, the F-measure results were generally very good (above 0.85) with the polynomial kernel. The only exceptions were the Bird



and Airplane (propeller) concepts, for which the linear and the radial basis functions kernel achieved the best results, respectively.

In order to evaluate these results, we built a small test set of documentaries, movies and TV shows. The results were not nearly as good. The worse performance can often be due to the fact that audio events almost never occur separately, being corrupted by music, speech and background noise. When analyzing the performance on movies, however, it is useful to not only look at F-measure results, but also take into account the very low number of positive examples for each concept in the whole movie. Hence, we also analysed the results of precision in terms of positive examples, $prp = tp / (tp + fn)$, and precision in terms of negative examples, $prn = tn / (tn + fp)$.

The concepts which achieved best results in this test set were Bird, and Telephone-Ringing (bell), for which we have achieved from 0.75 to 1.0 for prp and over 0.97 for prn . Intermediate results were obtained for Dog (barking), Airplane (propeller), and Crowd-Applause. For these concepts, some false positives were detected, resulting in a prp from 0.55 to 0.65 and around 0.95 for prn . For the Sirens and Water concepts, roughly the same amount of false positives were detected, but few true events were detected correctly. The concepts which achieved worse results were Airplane (jet), Gunshot, Helicopter, Horse (walking), Telephone-Ringing (digital) and Traffic, despite the very good results in the cross-validation set.

We are currently working on approaches based on hidden Markov models.

CERTH: Nothing to report during this period.

Task 3.2 Speech Recognition

INESC-ID: During the first 2 months of the 6-month period, the work of INESC-ID in terms of speech recognition consisted mainly of evaluating the performance of our automatic speech recognition (ASR) system for Broadcast News in US English with different corpora. These corpora were used in two evaluation campaigns promoted by DARPA/NIST. Our tests used the acoustic models trained on the basis of the HUB4'96

and HUB4'97 corpora. The results, in terms of Word Error Rate (WER), were the following:

Eval96 - 30.3%

Eval97 - 25.4%

Eval98 - 24.7%

Eval03 - 24.6%

As we had expected, the Eval96 corpus over which we had initially tested our ASR system is the most complex task of the four we have tested. Our results, namely for the most recent corpus (Eval03) are still far from the state of the art (12.3%), but the corpus material used for training in this evaluation campaign is not yet available. This corpus availability problem is hence of the major obstacles towards the continuing improvement of the ASR system.

Subsequently, we started working on a two-pass approach that takes into account the confidence measure of each word recognized during the first pass, and adapts the acoustic models of each speaker based on the words recognized with confidence higher than a threshold. This off-line procedure is currently performed using the information derived from the manual speaker segmentation, in order to evaluate the potential improvements due to the model adaptation without being affected by automatic speaker clustering errors. For the time being, this adaptation work was restricted to two corpora. Excluding the words with a confidence level lower than 0.955 excludes 25% of the targets for adaptation of acoustic models for RT03 and 29% for HUB'96. For the first evaluation corpus, the Word Error Rate improves to 23.3%. For the second evaluation corpus, it improves to 29.3%.

Our complementary work on topic identification and spoken language identification was also useful to the goals of this Work package.

CERTH: Nothing to report during this period.



- *Deviations*

None

- *Dissemination activities:*

INESC-ID:

Attended the VidiVideo meeting in Barcelona (Feb. 2008) and the 1st VidiVideo Review in Amsterdam (March 2008).

Relevant publications:

Hugo Meinedo, Márcio Viveiros, João Neto, Evaluation of a Live Broadcast News Subtitling System for Portuguese, INTERSPEECH 2008, Brisbane, Australia, September 2008.

Isabel Trancoso, José Portêlo, Miguel Bugalho, João Neto, António Serralheiro, Training audio events detectors with a sound effects corpus, INTERSPEECH 2008, Brisbane, Australia, September 2008.

Rui Amaral, Isabel Trancoso, Topic Segmentation in a Media Watch System, PROPOR 2008, Aveiro, Portugal, September 2008

Rui Amaral, Isabel Trancoso, Topic Segmentation and Indexation in a Media Watch System, INTERSPEECH 2008, Brisbane, Australia, September 2008

Jean-Luc Rouas, Isabel Trancoso, Céu Viana, Mónica Abreu, Language and Variety Verification on Broadcast News for Portuguese, Speech Communication, online publication complete: 10-JUL-2008, DOI information: 10.1016/j.specom.2008.05.006

CERTH:

Attended the VidiVideo meeting in Barcelona (Feb. 2008) and the 1st VidiVideo Review in Amsterdam (March 2008).

<i>Status of Deliverables by July 31, 2008</i>							
Deliverable	WP Leader	Status: completed / under way/not started yet		On schedule yes/no	Original completion date		Actual completion date calendar date
start date						1-feb-2007	
D3.1	INESC-ID	completed		Yes	month 12	1-feb-2008	1-feb-2008
D3.2	INESC-ID	completed		Yes	month 14	1-apr-2008	1-apr-2008
D3.3	INESC-ID	Underway		Yes	month 28	1-jun-2009	
D3.4	INESC-ID	Underway		Yes	month 26	1-apr-2009	

WP4: Visual Analysis: Workpackage leader: UvA
Co-reporting partners: UNIFI., CVC, UNIS, CERTH

- *Progress towards objectives:*

1: *Scientific & technical aspects*

The objective of WP4 was and still is to contribute to the visual part of the thesaurus. The goal is achieved by analyzing the context of images and videos characterizing objects and scenes. Over the past hal year, the aim was to achieve this characterization of objects and scenes by machine learning from invariant features derived from annotated examples. Features are derived by low-level analysis of static images and dynamic sequences and concept detection based on motion, human behaviour, frame composition, and visual scene information.

- *Work performed*

Task 4.1 Types of motion, actions & behaviour

According to the reviewer's comments, we first precisely define the Fuzzy Metric Temporal Horn Logic (FMTHL) developed last year. FMTHL was used last year as a rule-based inference engine in which conventional logic formalisms are extended by a temporal and a fuzzy component. This last one enables to cope with uncertain or partial information, by allowing variables to have degrees of truth or falsehood. The temporal component permits to represent and reason about propositions qualified in terms of time. These propositions are represented by means of conceptual predicates, whose validity is evaluated at each time-step. Next example shows the FMTHL inference rules for the concept `similar_direction`:

```
always(similar_direction(Agent, Agent2):-  
    has_status(Agent,_,_,_,Or1,_),  
    has_status(Agent2,_,_,_,Or2,_),  
    Dif1 is Or1 - Or2,
```

During these six months, all sources of knowledge from tracking were translated into this logic predicate formalism for subsequent reasoning. Motion trackers provide agent status vectors, which are converted into has status conceptual predicates:

$$t ! \text{ has status (Agent, X, Y, Theta, V)}$$

These predicates hold information for a global identification (instance id) of the agent (Agent), his spatial location in a ground-plane representation of the scenario (X, Y), and his instantaneous orientation (Theta) and velocity (V). A has status predicate is generated at each time-step for each detected agent.

Task 4.2 Types of scenes

Over the past 6 months, work have been done on a method for representing periodic events using spatio-temporal features. Within the PCA framework spatial nature of periodic events can be obtained by combining eigenvectors corresponding to periodic hidden variables.

In a uniform background, both object and background will have periodic behavior. Due to this the eigenvectors cannot be directly used as a spatial feature. When there is sufficient difference between intensities of object and background pixels, edges are expected to delineate object-



background boundary. Hence edge images combined with eigenvectors can be used as a spatial feature for representing motion events. The edges thus selected over a complete period of an event gives a spatio-temporal representation of the event. With such a representation, motion events can be considered as 3D point sets which can be compared using suitable dissimilarity measures.

Task 4.3 Types of objects

Our papers on the evaluation of color descriptors submitted to CVPR2008, CIVR2008 and CGIV2008 were accepted. We attended these conferences (June/July) to show our work through posters and presentations. During CIVR2008, we also co-organized the VideOlympics video search systems showcase. Also, our color descriptors were applied to the FRD cultural heritage data. Furthermore, we participated in the TRECVID2008 concept detection task, in co-operation with University of Surrey, who provided the multi kernel learning. This resulted in a joint submission to TRECVID2008.

- *Deviations*

None.

- *Dissemination activities*

Organization of the First International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS2008)

Organization of the VideOlympics video search systems showcase at the CIVR08

Participation in TRECVID08

Koen E. A. van de Sande, Theo Gevers and Cees G. M. Snoek, *Evaluation of Color Descriptors for Object and Scene Recognition*. Proceedings of CVPR. Anchorage, Alaska, USA, June 2008.

Koen E. A. van de Sande, Theo Gevers and Cees G. M. Snoek, *A Comparison of Color Features for Visual Concept Classification*. Proceedings of CIVR. Niagara Falls, Canada, July 2008.

R. Valenti and Th. Gevers, *Accurate Eye Center Location and Tracking Using Isophote Curvature*, IEEE CVPR, Alaska, USA, June 24-26, 2008.

A. Gijsenij, Th. Gevers and J. van de Weijer, *Edge Classification for Color Constancy*, IS&T's European Conference on Colour in Graphics, Imaging and Vision, Terrassa, Spain, June 9-13, 2008.

M. Lucassen, A. Gijsenij and Th. Gevers, *Comparing Objective and Subjective Performance Measures for Color Constancy*, IS&T's European Conference on Colour in Graphics, Imaging and Vision, Terrassa, Spain, June 9-13, 2008.

K. Sande Th. Gevers and C. Snoek, *Color Descriptors for Object Category Recognition*, IS&T's European Conference on Colour in Graphics, Imaging and Vision, Terrassa, Spain, June 9-13, 2008.



Ignasi Rius, Jordi Gonzàlez, Mikhail Mozerov, F. Xavier Roca, " Automatic Learning of 3D Pose Variability in Walking Performances for Gait Analysis ", International Journal for Computational Vision and Biomechanics, volume 1, number 1, pp. 33-43, January-June, 2008

Javier Varona, Jordi Gonzàlez, Ignasi Rius, Juan J. Villnueva, "On importance of detection for video surveillance applications", Optical Engineering, volume 47, issue 8, 087201, August, 2008

Carles Fernández, Pau Baiget, F. Xavier Roca, Jordi Gonzàlez, "Interpretation of Complex Situations in a Semantic-Based Surveillance Framework", Signal Processing: Image Communication, Special Issue on Semantic Analysis for Interactive Multimedia Services, volume 23, issue 7, pp. 554-569, August, 2008

Mikhail Mozerov, Ariel Amato, Xavier Roca, Jordi Gonzàlez, "Trajectory occlusion handling with multiple view distance minimisation clustering", Optical Engineering, volume 47, issue 4, 047202, April, 2008

Javier Orozco, F.Xavier Roca, Jordi Gonzàlez, "Real-Time Gaze Tracking With Appearance-Based Models", Machine Vision and Applications, doi: 10.1007/s00138-008-0130-6, available online April 4th, 2008

<i>Status of Deliverables by July 31, 2008</i>							
Deliverable	WP Leader	Status: completed/ under way/not started yet		On schedule yes/no	Original completion date		Actual completion date calendar date
start date						1-feb-2007	
D4.1	UvA	completed		yes	month 12	1-feb-2008	1-feb-2008
D4.2	UvA	completed		yes	month 14	1-apr-2008	1-apr-2008
D4.3	UvA			yes	month 28	1-jun-2009	
D4.4	UvA			yes	month 26	1-apr-2009	

WP5: Learning integrated feature software development:

Workpackage leader: UNIS

Co-reporting partners: UvA, Unifi

- *Progress towards objectives:*

1: Scientific & technical aspects;

We further investigate state of the art technique which already lead to significant improvement within task 5.2. The classification methods for the final system have been chosen, and we are performing extensive experiments to optimize the parameters and classification performance. We have adopted a software package for machine learning which will automatically fuse different features from WP4 and modalities which is the goal of task 5.3. Initial results for fused different feature types show further classification improvement.

D5.1 Reports on learning methods - Delivered

D5.2 Learning tools, first version - Delivered

Software library for learning classifiers from annotated image examples, with usage guidelines and performance analysis

D5.3 Learning tools, second version – in Progress

Software library for learning classifiers from annotated image examples, with usage guidelines and performance analysis

- *Work performed*

Task 5.1 One class active learning

Unis: In the past six months, our work has focused on two topics: object/scene classification with local features, and multiple kernel learning.

To improve the performance of one class classification we carried out the following experiments and doubled the performance compared to baseline results produced at the start of the project. We have now received trecdiv06 data containing 101 concepts and also trecvid07 data containing 36 concepts from the University of Amsterdam:

Recently local features have become popular in the field of object/scene classification. Typically, a set of local features (or a "bag of words", e.g. SIFT features that describe some regions, which are obtained either from keypoint detection, or from dense sampling) are used to represent an image. A similarity measure for two such representations is then defined. This similarity measure can be used as the kernel function in kernel-based learning algorithms such as an SVM. We experimented with two such bags-of-words based kernels: Pyramid Match Kernel (PMK), and Spatial Pyramid Match Kernel (SPMK). We used both the trecvid07devel data set (36 concepts, 18120 images in total) and the Mediamill data set (101 concepts, 43907 images in total). The table below shows the mean average precision (MAP) on the Mediamill test set using 4 different kernels:

sampling	feature	kernel	classifier	MAP
keypoint	sift	PMK	SVM	0.311
dense	sift	PMK	SVM	0.339
dense	colour hist.	PMK	SVM	0.252
dense	sift	SPMK	SVM	0.330

When looking at the average precision of individual concepts, we notice that different concepts have their own most suitable sampling/feature/kernel combinations. For example, for the "Snow" concept, dense sampling with colour histogram feature and PMK is the best; while for the "Airplane" concept, keypoint sampling with sift feature and SPMK is the best. This observation implies that information from different "channels" must be fused differently according to the concept being considered.

Different features can be of very different nature. For example, we may have one bag-of-words feature and one vector feature. In this case, it is not straightforward to fuse the information at the feature level. However, bag-of-words features and vector features can both be used to produce kernel matrices of the same size. It is therefore natural to fuse the information at the kernel level. One possible way of kernel-level fusion is multiple kernel SVM (mk-svm), where the goal is to find the "optimal" weights for the kernels.

We tried several multi-kernel machine learning toolboxes, and identified the Shogun toolbox was a good choice both in terms of fast convergence and good accuracy. We have managed to set up the toolbox and done some experiments. The table below shows the MAPs on the Mediamill test set using 3 kernels, the mean of the max performance on individual concepts, and the MAP of multiple kernel SVM using all 3 kernels.

sampling	feature	kernel	classifier	MAP
keypoint	sift	PMK	SVM	0.311
dense	sift	PMK	SVM	0.339
dense	colour hist.	PMK	SVM	0.252
max.				0.356
mk-svm				0.382

We can see from this table that mk-svm not only outperforms all kernels, but also outperforms the "max" of them. This is the case not only for MAP, but also for average precision of most concepts. This initial success shows that mk-svm is a promising approach for our problem. We are now running experiments using 4 kernels (this time including dense/sift/SPMK kernel), and will include even more kernels in the future, and test mk-svm on more data sets.

We also collaborated with the UvA, participating in the Trecvid 2008 challenge. We submitted a joint run using their kernels and our mk-svm.



Exploiting the multilabelness

=====

Semantic concept detection can be naturally formulated as a multilabel classification problem, where the semantic concepts a shot contains form its label set. To exploit this multilabelness, a binary classifier which we call the base classifier is trained for each concept. The base classifiers are then applied to the original feature vectors. Each feature vector is then augmented by concatenating the output of the base classifiers. Finally, another binary classifier is trained for each concept using the augmented feature vectors in a second round, to capture the potential dependency among the concepts. We extend this idea by replacing the binary output of the base classifiers with a probabilistic output provided by LibSVM, and by applying this process recursively.

This method is applied to two data sets: mediamill data (101 concepts, 120 dimensional visual feature provided by UvA, in total 120+101 dimensions) and trecvid07 data (36 concepts, $128 \times 16 = 2048$ dimensional region-based SIFT feature, in total 2048+36 dimensions). The method does improve over base classifiers. However, after a few iterations the performance tends to converge, or even starts to drop. Also, the improvement can be marginal when visual feature dimensionality is much higher than the number of concepts. Since in this scheme heterogeneous data are involved, the natural thing to try next is multi-kernel SVM as discussed in Task 5.2. We are also working on other state-of-the art classifiers as mentioned below.

Experiment1: Run basic classifiers (C4.5, Naïve Bayes, KNN, Logistic Regression) to find the precision on these concepts. The precision obtained from these classifiers will be used as base values for the state-of-art classifiers.

Experiment2: Run state-of-the art classifiers (Bagging, Boosting, Ensemble) to find the precision on these concepts and compared with base values. We also currently working on various over sampling methods e.g. SMOTE that generates synthetic minority (positive) class instances to solve class imbalance problem for many concepts. We are also investigating under-sampling methods.

Good results are obtained especially using Ensemble techniques and average mean precision is improved from 0.14 to 0.31. In the next few months; we will further investigate state-of-art methods to improve the precision.

In addition to that two main topics were investigated. i) The problem of effective processing of large training databases in AdaBoost was addressed. The developed approach, based on weighted sampling, reduces the available training set to much smaller active set on which the training is performed. The approximation is performed with respect to minimize the variance of hypothesis error estimate. We are now able to significantly speedup offline learning of object detectors. ii) The problem of online learning was addressed. Unlabeled examples in video provide a huge source of information that is in conventional detection systems not used. We would like to exploit such information in learning that seamlessly integrates a priori (offline) information with information gathered in tracking (online). We developed a preliminary version of object tracker based on boosting and learning in each frame. Our aim at this point is a system that is able to detect faces, track them and learn the face appearance; such approach would enable to track the face even in crowded scenarios.

Task 5.2 Learning semantic integrated feature detector

Pyramid Match Kernel (PMK): We applied PMK to the trecvid07devel dataset using three different features: keypoint/sift, region($2 \times 2 + 4 \times 4 + 8 \times 8 + 16 \times 16 = 340$)/sift, and region($2 \times 2 + 4 \times 4 + 8 \times 8 + 16 \times 16 = 340$)/color. The table below shows the mean average precision



(MAP) when using different features, where the structure of the vocabulary tree is fixed (8 levels, 16 nodes in each level):

feature	kernel	classifier	MAP
keypoint/sift b.o.f.	PMK (08/16)	SVM	0.2631
region(340)/sift b.o.f.	PMK (08/16)	SVM	0.3399
region(340)/color b.o.f	PMK (08/16)	SVM	0.2532

This table shows that region based sift descriptor outperforms the other two features. Moreover, when looking at the average precision of individual concepts, we notice that different concepts have their own most suitable feature. For example, for "Snow" concept, region based color histogram feature is the best; while for the "Airplane" concept, keypoint based sift feature is the best. This observation sits in well with the intuition that different concepts are of very different natures, and features must be combined differently according to the concept being considered.

Multiple Kernel SVM: One possible way of combining different features is to use kernel-level fusion in a multiple kernel learning (MKL) framework. This allows us to bypass the difficulty of combining bag-of-words feature and vector feature at the feature level. We are now experimenting with MKL and expecting to see some results soon.

UvA: Towards large-scale robust learning of integrated feature detectors, collecting a large amount of well-labeled representative visual examples is a crucial step. To solve the problem, we have started a study on exploiting existing image annotations in Flickr. Two key research challenges exist. First, the Flickr images are labeled by amateur users. The labeling is thus often very noisy. Our initial study on a subset of Flickr images shows that the labeling precision is around 49%. Existing learning algorithms can hardly handle such noisy training data. Second, since our target domain is video, how to adapt classifiers learned from image domain to video domain is an important problem. Based on the above observations, we propose a two-stage framework to solve the challenges. In the first stage, we improve Flickr tagging quality to gather good training examples. As an initial method, we propose a novel neighbor-voting algorithm to improve Flickr tagging quality. Afterwards, in the second stage, we will investigate multiple adaptive learning methods to overcome the domain diversity. We have visited the group of Josef Kittler at University of Surrey to discuss the possibilities for machine learning. We will verify our joint ideas in the concept detection task of the TRECVID 2008 benchmark evaluation.

- *Dissemination activities*

UNIS:

T. Tuytelaars and K. Mikolajczyk, "Local Invariant Feature Detectors - A Survey", Foundations and Trends in Computer Graphics and Vision, 2008.

K. Mikolajczyk and H. Uemura, "Action Recognition with Motion-Appearance Vocabulary Forest", In International Conference on Computer Vision and Pattern Recognition, Anchorage, USA, 2008.

H. Uemura, S. Ishikawa and K. Mikolajczyk, "Feature Tracking and Motion Compensation for Action Recognition", In British Machine Vision Conference, Leeds, UK, 2008.



F. Schubert and K. Mikolajczyk, "Combining High-Resolution Images With Low-Quality Videos", In British Machine Vision Conference, Leeds, UK, 2008.

Z. Kalal, J.G. Matas and K. Mikolajczyk, "Weighted Sampling for Large-Scale Boosting", In British Machine Vision Conference, Leeds, UK, 2008.

H. Cai, K. Mikolajczyk and J. Matas, "Learning Linear Discriminant Projections for Dimensionality Reduction of Image Descriptors", In British Machine Vision Conference, Leeds, UK, 2008.

<i>Status of Deliverables by July 31 2008</i>							
Deliverable	WP Leader	Status: completed/ under way/not started yet		On schedule yes/no	Original completion date		Actual completion date calender date
start date						1-feb-2007	
D5.1	UniS	completed		yes	month 12	1-feb-2008	1-feb-2008
D5.2	UniS	completed		yes	month 14	1-apr-2008	1-apr-2008
D5.3	UniS				month 28	1-jun-2009	

WP6: Technological software development:

Workpackageleader: CVC

Co-reporting: UvA, Unifi, Certh, Inesc-id, Unis

- *Progress towards objectives:*

- 1: *Scientific & technical aspects;*

- Definition of the syntax of files generated by the different software, using Mpeg7.
- Studying the possibility of creating open source software.
- A first integration phase has been finished where the file exchange formats have been defined.
- A second integration phase has been finished where all the pices of software are integrated using scripts
- Tests are being performed to the resulting software.
- Initial thesaurus of 572 detectors has been built.

- 2: *User- & market-related aspects*

None

- 3: *Management & partnership aspects*

- Communication via mail with all the partners in order to install/use/integrate their software.
- Communication via mail with all the partners in order to agree in a common exchange data format.

- *Work Performed*

Task 6.1 Software consolidation

After the last meeting, we contacted with all members of the project that develop software about the possibility of creating an open-source software. The answers where:

- UvA: Except form third part library, they use and create open source software. At present, they are using a non open-source library. However, this will change in the future.
- CVC: The PhD work is not open source until two years after the end of their Theses, but the software developed by engineers (like the GUI) is open source.
- Surrey: developing and using open source.
- Unifi: producing non open source software.
- Certh: possibility to make it open source
- Inesc-id: does not produce open source.

From the different answers we have arrived to the conclusion that it will be impossible to create full open source software because some of the members do not create open source. As a result, only a small subsystem will become open source.

In the first months we started to integrate the software developed by all the partners. The plan was -in this first phase- to have the software executed in different machines (each of the partners will execute his own software), but using the output of another software as the input. This was a first step towards integration. When this step finished correctly, we would proceed into a more

in-depth integration. The most important aspect of this phase was to establish the syntax of the input/output files.

Before starting this process, we defined a document where MPEG7 was defined in the Activity Report of the first year of VIDIVideo. In this document we described, for each output/input of the software, how the MPEG7 files should be (which information has to store and how). This document (with the modifications) can be found at [surfgroepen.nl](http://www.surfgroepen.nl/sites/vidi-video/Shared%20Documents/WP6%20Technological%20software%20development/Mpeg7Descriptions.doc) in:

[shared_documents/Wp6/Mpeg7Descriptions.](http://www.surfgroepen.nl/sites/vidi-video/Shared%20Documents/WP6%20Technological%20software%20development/Mpeg7Descriptions.doc)

<https://www.surfgroepen.nl/sites/vidi-video/Shared%20Documents/WP6%20Technological%20software%20development/Mpeg7Descriptions.doc>

In brief the process consists of:

- **Visual Segmentation:** This process is done by CERTH. This process receives as input a video file. It outputs an Mpeg7 with the video shots. The content of the Mpeg7 is defined in the document.
- **Audio segmentation:** This process is done by Inesc-id. This process receives as an input a video file. It outputs an Mpeg7 with the audio shots.
- **Visual Features:** This process is done by UvA. This process receives as an input a video and the visual shots defined in the Visual Segmentation. It outputs a file with the visual features of the keyframes of the shots.
- **Learning / classification:** this process is done by Surrey. This process has two functionalities. The first one is, receiving the visual features and video annotations, create classifiers (this are stored in an internal format, not Mpeg7). The second functionality is, given visual features use the classifiers to create annotations
- **GUI:** We have tested the CVC Gui for this experiment. Given the annotations of videos, the GUI will enable the user to search into the videos.

During these executions, the Mpeg7 format was defined and modified in order to fulfill the software requirements. At the moment the format is completely defined and, for the moment, will not be changed.

It is important to note that CVC has not investigated whereas the outputted results were ok (their quality). The unique interest was in defining the file exchange format.

When the first phase was finished CVC started a second integration phase. The aim of the second phase was to create a single piece of software which integrates all the different elements. We started by recollecting the different software used in the previous phase and installed it in our local machines.

We have been working on installing and executing the different software. Moreover we have developed several software/scripts in order to install/execute all the software with a single command. However we have found some problems (that we could solve with the appreciated help of all the partners) and an inconvenient one: Certh software has to be run in Windows whereas other software runs on Linux only.

After some hard work we manage to create two pieces of software one for Windows and the other one for Linux. The windows software enables to create the segmentation files for videos. The Linux software enables to do the rest of the work. The resulting software has a considerable



size so we have decided to store it in one of the CVC servers so it can be accessed by every one in the project. The files and links are the following:

The readme file:

<http://ise.cvc.uab.es/mateu/vidivideoSoftware/software/readme.pdf>

Windows software (as described in the readme file):

<http://ise.cvc.uab.es/mateu/vidivideoSoftware/software/windowsSoftware0.1.rar>

Linux software (as described in the readme file):

<http://ise.cvc.uab.es/mateu/vidivideoSoftware/software/linuxSoftware0.1.tar.gz>

Briefly we can say that to run the software the user will need to (however all this information is described extensively at the readme file stated above):

1. We need a windows machine, and an Ubuntu 7.10 'new' (with the ubuntu just installed, no other software) machine.
2. Install the software
 - a. Uncompress a .rar file in Windows
 - b. Uncompress a .rar file in Linux, and run an installation script
3. Learning phase
 - a. Run a script in Linux
4. Classification phase
 - a. Run a script in Windows
 - b. Send the data to Linux machine
 - c. Run a script in Linux

Due to the intrinsic difficulty of integrating software from different partners, we had no time to make tests to the software. Moreover, the execution time required is quite big (as an example, the visual segmentation of the videos has taken about 2 weeks to compute), and therefore the time of doing test is also big. At the current time we are executing the software to a big set of videos to test its correctness.

For these reasons we do not expect to finish the test until September.

UvA: nothing to report

Inesc-id: nothing to report.

Unifi: nothing to report

Certh: nothing to report

Task 6.2 Thesaurus of detectors

UvA:

We have made an initial thesaurus of 572 detectors which have been applied to TRECVID 2004, TRECVID 2005, TRECVID 2006 and TRECVID 2007 data. The performance varies widely from being good to close to random.

Performance depends mainly on the quality of the annotations and whether the annotations are representative for the dataset on which they are applied.

CVC: not involved in this task

Unifi: not involved in this task

Certh: not involved in this task

Inesc-id: not involved in this task

Task 6.3 Benchmark evaluation

UvA: nothing to report.

CVC: not involved in this task.

Unifi: not involved in this task

Certh: not involved in this task

Inesc-id: not involved in this task

- *Deviations*

The initial plan was that the first software was finished on the 10th June. However, there was one week delay due to the difficulty of running and integrating software produced in different universities.

- *Dissemination activities:*

CVC: nothing to report

UvA: nothing to report

Unifi: nothing to report

Certh: nothing to report

Inesc-id: nothing to report

<i>Status of Deliverables by July 31, 2008</i>						
Deliverable	WP Leader	Status: completed / under way/not started yet	On schedule yes/no	Original completion date		Actual completion date calendar date
start date					1-feb-2007	
D6.1	CVC	completed	yes	month 12	1-feb-2008	1-feb-2008
D6.2	CVC	completed	yes	month 16	1-jun-2008	1-jun-2008
D6.3	CVC	Underway		month 30	1-aug-2009	
D6.4	CVC	Underway		month 21	1-apr-2009	
D6.5	CVC	Not started yet		month 35	1-jan-2010	

WP7: Demonstrator and application: Workpackage leader: UNIFI
Co-reporting partners: CVC, UvA, B&G and FRD

- *Progress towards objectives:*

- 1: *Scientific & technical aspects;*

The scientific aspects of the WP are related to the development of ontologies that are suitable for the task of video annotation and retrieval. Ontologies for the three video domains used within VidiVideo have been defined.

A paper describing a methodology to include multimedia information in Web Ontology Language (OWL) ontologies has been accepted for publication in IEEE Multimedia, and a schema of the OWL ontology format has been published on the VidiVideo web site.

The ontology developed for documentaries has been used within the task of creation of manual annotation of documentary/cultural heritage videos, that aims at providing a large set of audio-visual concept samples that can be used within WP5, for training of concept detectors. The tool employed to create the manual annotation has been developed within the WP.

An algorithm that learns Semantic Web Rule Language (SWRL) rules, that can be used to detect complex concepts has been developed. The technique is based on the First-Order Inductive Learner (FOIL), and allows to learn first-order logic rules.

The technical aspects are related to the software development of tools for the creation of manual annotation, and browsing/querying annotated videos for the three different video domains. The tools that have been developed follow the guidelines reported in the deliverable D7.2.

- 2: *User- & market-related aspects*

Presentations of automatic and assisted video annotation to content providers have been carried on, resulting in the interest of the Italian national broadcaster RAI, that is particularly interested in the development of concept detectors.

- 3: *Management & partnership aspects*

UNIFI, CVC and UvA have collaborated on the development of D7.2. UNIFI and B&G have collaborated during the development and use of the manual annotator tool. UNIFI and FRD have organized the presentation of the VidiVideo project to the Italian RAI national broadcaster R&D department.

UvA has provided FRD with a standalone GUI program that can be used on portable computers during dissemination meetings and demos.

- *Work Performed*

Task 7.1 Ontology of queries

UNIFI has worked on methods for the automatic learning of rules that describe events, based on the use of the Web Ontology language (OWL) to describe the video domain of interest, and the Semantic Web Rule Language (SWRL) to model the rules. The goal is to overcome the problem of the necessity of a human expert that has to model the rules for a certain domain: this approach is not practical for the definition of a large set of rules. At present the majority of the solutions proposed in scientific literature is based on methods that learn a set of rules by exploiting decision tree algorithms and low-level features, or simple junctions of high level concepts, and are not enough expressive to describe complex events. For example consider the event *A person enters in secured area*, typical of a video surveillance video. This event can not be described

using only the low-level descriptors of the person and of the area, or using the co-occurrence of the high-level concepts *person* and *secured area* since the person may stay outside of it, or may have always been inside it; instead it is required to take into account the temporal evolution of the characteristics and features of the objects and entities. In fact, this event can be fully described and modelled using first-order logic (FOL). The event that would be described using the following sentence:

IF a person is outside of the secured area in the interval t_1 AND the same person is in the secured area in the interval t_2 AND t_1 is before to t_2 THEN that person has entered the secured area.

It can be translated in the following fragment of first-order logic language:

IF $\text{person}(p) \wedge \text{personOutsideOfSecuredArea}(p, t_1) \wedge \text{personIsInSecuredArea}(p, t_2) \wedge \text{before}(t_1, t_2)$
THEN $\text{personEntersSecuredArea}(p)$

where p is a variable that can be bound to any person and t_1 and t_2 are variables that are used to represent time intervals.

A method to learn sets of first-order logic rules applicable directly to the standard SWRL and OWL ontology for the description of events has been devised. This method is an adaptation of First Order Inductive Learner technique (FOIL, proposed by Quinlan) to the Semantic Web technologies and in particular to the automatic definition of events; for convenience this method will be referenced in the following as FOILS.

This approach permits to create an ontology structure that allows to perform automatic semantic annotation of video sequences matching visual descriptors and recognizing events described using rules, that have been automatically learned. Moreover the learning approach used is more expressive than the methods based on decision trees and concept conjunctions, because it defines rules through the first order logic theory.

The hypotheses learned by FOILS, similarly to FOIL, are sets of FOL rules, where each rule is similar to a Horn clause with the limitation that literals are not permitted to contain function symbols, in order to reduce the complexity of the hypothesis space search.

This approach has been tested to learn rules that describe airplane events described in the LSCOM ontology, using the LSCOM 2005 development set videos, and on surveillance videos selected from the public CAVIAR video set. A paper describing the approach has been accepted in a scientific conference (see Dissemination section).

A keynote presentation including this development will be presented at the ACM International Workshop on the Many Faces of Multimedia Semantics.

A prototypical system that creates relations using WordNet and the analysis of co-occurrence of concepts in a video is being developed. The goal is to test its application for both extending the automatic annotation of videos and as a tool that can be used in multimedia query tools, to let users see the connections between concepts, and allow to automatically extend a user query using these relations. The tool is being tested also to improve the recognition rate of automatically detected concepts by changing the confidence thresholds of the detectors, based on the co-occurrence of concepts and on the mutual information that can be determined by this.

UvA has contributed new concepts to the ontologies of queries, starting from the past Amsterdam meeting.

UNIFI and B&G have extended the lexicon of the documentaries ontology since the initial use of the manual annotator tool (see task 7.3).



Task 7.2 Multimedia query tool

UNIFI has worked on the development of the query tool for documentaries/cultural heritage. Following the schema outlined in deliverable D7.2 this application is based on the RIA paradigm. Users can query annotations stored in an ontology (expressed using OWL) to build queries within a browser, and the corresponding video sequences are streamed through the internet. In its initial form the tool uses the annotations created by B&G (see task 7.5). This approach will ease the task of presenting the system to interested users such as video archive institutions dealing with cultural heritage, since videos belonging to this video domain have been used.

CVC has been working on a system to store instances of an ontology into a database. This software is being programmed using Java, and using a MySQL database. However, it is planned to be functional for any SQL database (due to the use of Hibernate object/relational persistence and query service, that maps the Java classes to SQL database tables).

The first objective of the software is to receive an input with some instances of events, interpret the events with a given ontology, and store these events with the information retrieved from the ontology to the database. The second objective is to create complex queries to obtain information from this database.

The software is able to, given a file with events generated by the CVC annotations tools, retrieve semantic data from a given OWL ontology, and store the data to the data base in a comprehensive format.

UvA, has prepared a new demonstrator GUI that can be used to perform queries on some pre-processed test videos. The demonstrator can be executed on a portable PC; so to ease the demonstration activity of FRD. The videos have been processed with the automatic concept detectors developed within WP4 and 5.

Task 7.3 Interaction and visualisation

UNIFI has developed a manual annotation tool, using the RIA paradigm described in D7.1 and D7.2. The tool allows to annotate video sequences using the ontology developed in the past months or adding new ontologies created by users. The goal of the tool is to create ground-truth annotation that can be used during the training phases required to build automatic concept detectors in WP5. The tool is also useful to drive the design of user interfaces, since it has started to be heavily used by B&G annotators, that are contributing their comments and suggestions on the different components of the interface. The deployment of the tool is providing also valuable insights regarding the problems that can be encountered during a heavy use of RIA technologies for video applications. Some components developed for this tool have been re-used for the RIA query tool.

CVC has been working on the Broadcast/News GUI. The first main goal has been working in increasing the efficiency of the GUI. This is a very important goal, because it is one of the requirements of the software. CVC has applied software strategies to increase the efficiency. Moreover CVC has solved a problem that was encountered with a version of OpenGL and ATI graphic cards. The second goal has been the increase of the usability of the software. CVC has started working on increasing the interactivity between the user and the application in order to create a more intuitive interaction, while trying to reduce the amount of time needed to search videos (that, as has been told, is the main goal of this software). Finally CVC has been working on preparing the application to be exported outside the CVC software build system.

The main features of the Broadcast/News GUI are:

- Search into the data by introducing the desired keyword.
 - The display of the results is done by a 3d interface.
 - Two different representations have been designed in order to adapt to the different users.
 - The way that the GUI reacts has been done in order to create a user-friendly interface.
-
- Using mouse events navigate through the search.
 - Some advanced search features have been designed.



- The user may select in which videos he wants to search.
- Multiple database control has been designed – the user can select which of the databases he wants to search in.
- Navigate through the videos stored in the database.
- Inspect the different shots.
- Inspect the data which has been obtained in the different shots.

CVC has also worked on the integration of the GUI with the rest of the software as explained in WP6 deliverable, creating some shell applications in order to introduce data to the application. The application has been successfully ported to Linux.

Finally CVC has continued working on introducing natural language queries to the GUI. This work is still in its early stages.

Task 7.4 User groups

During the last six months FRD has produced a questionnaire in order to define the users' requirements (in particular for the cultural heritage/scientific field and for the broadcasting) focusing on the subjects and institutions to be involved in the definition of the users' requirements and also published it on the web, for two months till the end of July, at the following URL:

www.surveymonkey.com/s.aspx?sm=RWaVC7S5kTV75upJOXy2FQ_3d_3d.

FRD shared the questionnaire with UniMoRe (UNIFI subcontractor) that adapted it for the videosurveillance field and submitted it to a number of possible users during a meeting (22nd of May, Modena).

FRD has submitted the questionnaire along with an interview among a selected group of Italian Video Archives, in the Cultural Heritage and scientific field in order to take closest contacts for the next step (field trials) once we have the prototype of videos products (documentaries and short films). and to have new videos and materials to work on for the prototype (Mediateca Regionale Toscana, Cineteca Nazionale, Roma, AAMOD, Festival dei Popoli Firenze, Museo del Cinema Torino, Cineteca di Bologna). This was important not only to involve and make aware these Institutions about the project, but also to understand the state of the art on AV cataloguing and researching matters and best practises. FRD has asked the collaboration of the partners in order to enlarge as much as possible the submission of the questionnaire in each country and video domain and also to create a list of possible users. A number of responses were received, and FRD is working on the analysis of the questionnaires results for the production of the users' requirements document to be submitted to the technical partners.

The interviews, on one hand revealed great interest in the future achievements of the project and of the potential features of the software, on the other hand showed a technological gap between the basic technologies currently used by most of these institutions and the solutions proposed by VidiVideo. In fact the technologies developed and integrated by VidiVideo are not immediately suitable for a large part of the end users. For these reasons a special effort will be taken to allow users to be able to fully understand and test the software solutions and by consequence be able to effectively contribute to the definition of the final users requirements.

Task 7.5 Cultural heritage documentaries

UvA, has provided a demo system to FRD, showing the annotation results of the videos provided by FRD in the past months. The system will be used to attract more interest from possible users in the cultural heritage domain.

UvA, has processed several documentaries videos, to use the automatic annotations for a demonstrator of the query system.

B&G has started the annotation of a new corpus of documentaries videos, to provide ground truth annotation of the concepts that were defined in the documentaries ontology. A person has been hired specifically for this task and works 40 hours/week on the project. The tool that is being



used is the one developed by UNIFI (see task 7.3), and is based on Rich Internet Application (RIA) paradigm. The annotator has added several new concepts to the documentaries/cultural heritage ontology, while reviewing and annotating the videos. At present more than 14.000 annotated concepts have been produced. This effort will help to train concept detectors that are specific for documentaries videos, with the goal of improving the recognition performance of audio/visual concepts that are specific for this domain. This will have an impact on the future results of Task 7.4, since at present the majority of concept detectors have been trained to recognize concepts related to the news video domain. The intensive annotation work performed by B&G has helped UNIFI to recognize and solve various issues related to the development of RIA video tools, that have been beneficial for the development of RIA query tools developed in D7.3.

Task 7.6 Broadcast archive field trial
N/A

Task 7.7 Video surveillance

UNIFI sub-contractor (UniMoRe) has continued development of the video repository, adding functionalities for MPEG-7 annotation export.

CVC and UniMoRe have added some new videos to the video surveillance repository.

UNIFI has produced two “car” detectors (one for front and one for side views), based on Haar cascades, that will be distributed through the VidiVideo collaborative web site. These detectors have been developed since they are particularly important for the recognition of surveillance events.

UNIFI has continued research activity for video surveillance; in particular the activity has regarded the problem of estimating on-line the time-variant transformation relating a person’s feet position in the image of a first, fixed camera, to his head position in the image of a second, pan-tilt-zoom camera. The transformation allows acquiring high-resolution images by steering the PTZ camera at targets detected in a fixed camera view. Assuming a planar scene and modelling humans as vertical segments, an uncalibrated framework which does not require any 3D known location to be specified has been developed, and it allows to take into account both zooming camera and target uncertainties. Results show good performances in slave camera target head localization, degrading when the high zoom factor causes a lack of feature points in the slave camera.

Another key issue in surveillance video analysis is the problem of tracking a moving target. For this task a novel method for tracking a human target based on uncalibrated video analysis techniques acquired by rotating and zooming camera sensors is being studied. In these sequences targets may appear and disappear from the field of view (FOV) and their size may change very much in a few frames as a result of the camera zoom or target motion. Moreover targets generally change their size when they exit and then re-enter the FOV of the camera.

By approximating targets as vertical sticks moving on a planar surface it is possible to decouple target imaged size from its position and speed so that a simplified target state model can be used that includes only the target location. This simplifies the measurement extraction process when a target reappears in the FOV.

UNIFI is working on multitarget tracking, and is studying the use of ontologies for the detection and recognition of surveillance video events.

- *Deviations*

UNIFI has recruited three persons, that starting from March will replace the researchers that left the group.



- *Dissemination activities:*

UNIFI has submitted the following papers (accepted):

- Alberto Del Bimbo, Federico Pernici “Uncalibrated 3D Human Tracking With A Ptz-Camera Viewing A Plane” Proc. 3DTV International Conference: Capture, Transmission and Display of 3D Video (3DTV-CON 08), Istanbul, May 2008.
- A. Del Bimbo, F. Dini, A. Grifoni, F. Pernici “Uncalibrated framework for On-line Camera cooperation to acquire human head imagery in wide areas”. 5th IEEE International Conference On Advanced Video and Signal Based Surveillance Santa Fe, New Mexico September 1-3, 2008
- M. Bertini, A. Del Bimbo, G. Serra, “Learning rules for Semantic Video Event Annotation”, Proc. 10th International Conference on Visual Information Systems (VISUAL) 2008
- M. Bertini, A. Del Bimbo, G. Serra, “Learning Ontology Rules for Semantic Video Annotation”, keynote paper at the 2nd ACM International Workshop on the Many Faces of Multimedia Semantics, Vancouver, October 2008

UNIFI and the subcontractor UNIMORE submitted a journal paper (accepted) to IEEE Multimedia describing the multimedia ontology framework studied within task 7.1. Date of publication to be determined.

- M. Bertini, R. Cucchiara, A. Del Bimbo, C. Grana, G. Serra, C. Torniai, R. Vezzani, “Dynamic Pictorially Enriched Ontologies for Video Digital Libraries”

The UNIFI subcontractor UNIMORE submitted two papers (accepted) on the Video Surveillance repository developed within the project:

- R. Vezzani, S. Calderara, P. Piccinini, R. Cucchiara, "Smoke detection in videosurveillance: the use of VISOR (Video Surveillance On-line Repository)" in press on Proceeding of ACM International Conference on Image and Video Retrieval, Niagara Falls, Canada, July, 7-9, 2008
- R. Vezzani, R. Cucchiara, "ViSOR: Video Surveillance On-line Repository for Annotation Retrieval" in press on Proceedings of IEEE International Conference on Multimedia & Expo (IEEE ICME 2008), Hannover, 2008

UNIFI and FRD have presented the VidiVideo project, and some of its deliverables to the Italian national broadcaster RAI R&D department, that is responsible for the development of the annotation tools currently being used by the RAI archives (July 21st, 2008).

N.B. the deliverables are numbered according to page 60 of the DoW!

<i>Status of Deliverables by July 31 2008</i>							
Deliverable	WP Leader	Status:completed/ under way/not started yet		On schedule yes/no	Original completion date		Actual completion date calendar date
start date						1-feb-2007	
D 7.1	UNIFI	completed		yes	month 12	1-feb-2008	1-feb-2008
D 7.5	UNIFI	completed		yes	month 12	1-feb-2008	1-feb-2008
D7.7	UNIFI	completed		yes	month 12	1-feb-2008	1-feb-2008
D7.2	UNIFI	completed		yes	month 14	1-apr-2008	1-apr-2008
D7.3	UNIFI	complete d		yes	month 18	1-aug-2008	17-aug- 2008
D7.4	UNIFI				month 32	3-okt-2009	
D7.6	UNIFI				month 34	1-dec-2009	
D7.8	UNIFI				month 32	1-okt-2009	
D7.9	UNIFI				month 32	2-okt-2009	
D7.10	UNIFI				month 32	3-okt-2009	

WP8: Data, annotation and queries:

Workpackage leader: B&G

Co-Reporting partners: Unifi, Certh, UvA

- *Progress towards objectives:*

1: *Scientific & technical aspects;*

1. Audio event detection

In collaboration between partners B&G and INESC-ID, a plan was drafted regarding audio concept detection. As a first step, B&G supplied an export of metadata from their in-house sound effects collection. (approximately 6.000 cds). INESC-ID distilled a list of audio concepts from this collection. In a next phase, researchers of INESC-ID visited B&G April to discuss next steps. 280 cd's have been copied in a hard-drive and where sent to INESC-ID early may. This data set (including the metadata descriptions) will be used for audio concept detection.

2. Manual annotations

For testing purposes, a selection of the total video collection will be annotated on a more fine-grained level, using the list defined by the project. B&G selected a five hours corpus especially for visual event annotation.

From the data set, five hours are manually annotated with a selected subset of concepts from the preliminary list based on [1] Relevance in the documentaries [2] The number of occurrences in the dataset. [3] Diversity in the type of concept. A description of this collection can be found in D8.3.

3. Co-design of the annotation tool

UNIFI created a special tool for the concept annotation of the videos, for both the Audio and Video concepts. Using this tool it is possible to annotate video sequences associating video segments with concepts selected from different ontologies. The tool was released mid-May 2008. B&G tested this tool and provided input for improvements during May-August, so the tool was continuously updated.

From the third week of May 2008 annotations of the 5 hours of video content have been created at at B&G this will be finalized by September.

4. Development of the Query Log environment

As part of task 8.3 Sound and Vision is developing a Query Log environment, that will be directly linked to the iMMix catalogue. In April, this tool has been installed and tested. In May, the results fo this activity was discussed with the software developers.

D8.1 500 hours of digital video

D8.2 Annotations

D8.3 Annotations for other relevant descriptions

D8.4 Query logs and the analysis

- *Work Performed*

Task 8.1 Data, annotation & queries

B&G has provided test data (cd's with audio events) and is working on manual annotations. See also 'work done' section above.

Unifi finalised the software required for annotation.

UvA helped defining the dimensions of the test collection.

Task 8.1 Query Log Analysis

B&G conducted initial query logs analysis based on a fixed set of query log data. Vidi-Video however aims to analyse query behaviour in a more structural fashion. For this task, a specialised module needed to be built. This was done in earlier reporting periods and this tool has been installed and tested in April. In May, the results of this activity were discussed with the software developers. It is expected to launch the Query Log tool in August 2008 (see below).

- *Deviations*

The query log module from B&G has become operational later than anticipated. Major reason for this is the interdependency of the in-house software development roadmap at B&G. The external software company are working on the development of the software development. They are responsible for the timely of system components that are critical to the B&G operations. Delays in the development of these components in the first quarter of 2008 caused a delay also in the development of the query log analysis tool, built specifically for Vidi-Video. It was impossible to influence this process.

The development cycle of this module took up much more time than could be anticipated earlier. Originally, the first results should have been documented in PM16 (June 2008). Given the delays mentioned above, it is now estimated that first results will not become available until October, as the tool will become available in August and needs to build up a database of queries to be analysed.

As a countermeasure, B&G managed to set up smaller scale analyses (based on static logs) and these results have been shared with the consortium. In the DoW, it is stated "These query logs will be analyzed to derive the set of most relevant concepts which need to be in the thesaurus.". The results of these small scale experiments provided input for defining the visual thesaurus. The impact of this delay on the overall progress of the Vidi-Video system development can be labelled as low.

- *Dissemination activities:*

B&G organized a two-day Archive Seminar "Employing Values" on the values of audiovisual collections and how to exploit those, jointly organised by Beeld & Geluid, EBU and FIAT/IFTA on 6-7 March 2008. Vidi-Video results were presented here.

Full report:

http://fiatifta.org/conferences/seminars/past/hilversum_2008/SeminarReport_FINAL.pdf

B&G organized a three day international conference in De Balie in Amsterdam on 11 and 12 April 2008 on the topic of online access to audiovisual heritage. The Vidi-Video project was presented as part of the first plenary session.

UvA and **B&G** hosted a one day academic event on June 27th in Hilversum Around 50 researchers were present. One of the presentations was made by Prof. Dr. Arnold Smeulders on the Vidi-Video project and related research.



Planned activity:

- September IBC 2008 www.ibc.org
- September FIAT/IFTA www.fiatifta.org
- November TRECVID WORKSHOP <http://www.nist.gov>

N.B. the deliverables are numbered according to page 60 of the DoW!

<i>Status of Deliverables by July 31, 2008</i>							
Deliverable	WP Leader	Status: completed / under way/not started yet		On schedule yes/no	Original completion date		Actual completion date calendar date
start date						1-feb-2007	
D8.2	B&G	completed		yes	month 3	2-mei-2007	2-mei-2007
D8.3	B&G	completed		yes	month 15	1-mei-2008	1-mei-2008
D8.4	B&G				month 30	1-aug-2009	
D8.5	B&G			No, ready in October (see above)	month 16	1-jun-2008	

WP9: Dissemination: Workpackage leader: **FRD**
Co-reporting partners: CVC, UvA

- *Progress towards objectives:*

1: *Scientific & technical aspects;*
Not relevant.

2: *User- & market-related aspects*

New contacts were taken with other Italian Cultural Heritage institutions operating in the digital audiovideo sector in order to involve them in the interviews and in the setting up of the user groups.

The questionnaire created in order to define the User requirements (see WP7) has been submitted to a group of selected Italian Cultural Heritage institutions operating in the digital audiovideo sector, and in particular: Mediateca Regionale Toscana, in Florence; Cineteca Nazionale, Rome; Festival dei Popoli Florence; Museo del Cinema Turin; Rai, Turin; Cineteca di Bologna; Istituto Luce, Rome: most of them have agreed to test the prototype once it will be ready and will give support to the project by providing new audiovisual material for testing. The partners have provided a number of new contacts for the submission of the questionnaire.

On the side of broadcasting FRD has also directly taken contacts for the submission of the questionnaire and for giving awareness to the project and its further developments and achievements with BBC London; ORS (Austrian Radio Broadcasting), National Greek Archives; MemoriaV, (Swiss association for audiovisual preservation), TVCataluna; Deutsche Film Institute, Deutsche Welle Germany's International Broadcasting Station, SLBA (Swedish national Broadcasting Station).

Following up the great amount of suggestions and information given by the partners on meetings, workshops and events FRD is studying case by case and taking further information in order to organize and focus on the most important of them in order to provide a selected list of proposals to be chosen by the partners.

3: *Management & partnership aspects*

The activities of community building are proceeding mainly with other cultural Italian institutions and organizations, but now thanks to the interview-questionnaires some good contacts in other countries area arise. The activities of community building are proceeding with other cultural Italian institutions and organizations and all the partners are working in extending the interviews in their own countries.

On the same matter an answer of the main EC founded project is awaited.

- *Work Performed*

Task 9.1: Public awareness and dissemination

Task 9.1 – Public awareness and dissemination

The partners of the project have approved the Web site of the project and now the site is on line at the address www.vidivideo.eu and www.vidivideo.info.

The plan of use and disseminating knowledge has been validated by the partners.

Different demonstrations of interest have been collected from different Italian cultural institutions with significant audiovisual archives.

After the good results obtained during the interview-questionnaires phase and declaration of potential interest by some participants in the user group on culture, the next step for dissemination depends by the availability of a ‘convincing’ prototype in order to develop the initial work with experts in a more concrete and refined user requirements and validation of the technology under production by VIDIVIDEO.

Task 9.2 Networking and Clustering of the project

Contacts have been established with the leaders of “The production of content information (metadata) allowing for access and delivery “ work package of the EC project PRESTOSPACE (<http://www.prestospace.org/>), to verify how to start a common activity.

Contacts exist also with the MULTIMATCH EC-funded project (www.multimatch.org) and we discuss the possibility to organise jointly an event to raise awareness and demonstrate potential use of advanced technologies for AV archives.

We’ve also contacted the community of the DELOS Network of Excellence (<http://www.delos.info/>) to define how to involve the DELOS community in the VIDI-Video project.

FRD has organised a thematic workshop on music archive “Workshop on digitisation and long-term preservation of sound archives” on 15-17 November 2008 in cooperation with AXMEDIS, DELOS and many relevant institutions or individual experts in the field. Even if the topic is more specifically on videos, most of the issues are the same or strictly related and, more important, the target community is in part overlapping, so it is expected a good impact and cooperation possibility from this event.

- *Deviations*

The workpackage is proceeding as scheduled.

- *Dissemination activities:*

The activities of community building are going on, and personnel of FRD have participated at different meetings with institutions that can be involved in the project.

The first annual report of the project, the “*VIDI-Video Annual Report 2007*”, have been produced and delivered to the European Commission, who will provide online publishing.

The UvA has organized the 2nd VideOlympics as part of the CIVR 2008 in Niagara Falls (see www.videolympics.org). The VideOlympics is a video search showcase where systems are simultaneously competing on an interactive search task.

Organizers:

Cees Snoek, *University of Amsterdam, The Netherlands*

Marcel Worring, *University of Amsterdam, The Netherlands*

Rong Yan, *IBM Research, USA*

Alex Hauptmann, *Carnegie Mellon University, USA*

Co-organizers:

Ork de Rooij, *University of Amsterdam, The Netherlands*

Koen van de Sande, *University of Amsterdam, The Netherlands*



Invited presentation:

M. Worring et.al., Semantic search and image detector technology, Video Search Summit, San Francisco 2008, an event with participation of commercial companies like ClipBlast, Truveo, and CastTV.

<i>Status of Deliverables, by July 31, 2008</i>							
Deliverable	WP Leader	Status:completed / under way/not started yet		On schedule yes/no	Original completion date		Actual completion date calender date
start date						1-feb-2007	
D 9.1	FRD	Completed		Yes	month 12	1-feb-2008	1-feb-2008
D9.5	FRD	Completed		Yes	month 12	1-feb-2008	1-feb-2008
D9.6	FRD	Completed		Yes	month 12	1-feb-2008	1-feb-2008
D9.2	FRD				month 26	1-apr-2009	
D9.3	FRD				month 28	1-jun-2009	
D9.4	FRD				month 35	1-jan-2010	

2. Project Efforts

Period 7, February March 2008

Period 7, month 13 and 14

Part. Short name	WP1	WP2	WP3	WP4	WP5	WP6	WP7	WP8	WP9	Total	Old total	This period's total	New actual total	Planned Total	Deviation	Comments on deviations
UvA management	2.96									2.96	7.8	2.96	10.76	9.1	18%	
UvA RTD				7						7	21.5	7	28.5	22.69	26%	
CERTH/HIT		1.71								1.71	7.03	1.71	8.74	9.48	-8%	
INESC-ID		1.4	2.85			0.1				4.35	11.12	4.35	15.47	15.28	1%	
UNIS				0	1	0				1	19	1	20	15.48	29%	
UNIFI-MICC							2			2	25.5	2	27.5	18.76	47%	
CVC				1		2	1			4	23	4	27	12.09	123%	This deviation is due to our initial efforts on integrating the modules developed at CVC in WP4 and WP7. This has allowed us to know how to integrate the modules of the rest of the partners for the rest of WPs
BandG								1		1	7	1	8	5.77	39%	
FRD							0.1		0.1	0.2	5.15	0.2	5.35	5.33	0%	

Period 8, April May 2008

Period 8, month 15 and 16

Part. Short name	WP1	WP2	WP3	WP4	WP5	WP6	WP7	WP8	WP9	Total	Old total	This period's total	New actual total	Planned Total	Deviation	Comment on person months
UvA management	1.69									1.69	10.76	1.69	12.45	10.4	20%	Please comment
UvA RTD				6.35	1.07	2.71				10.13	28.5	10.13	38.63	29.97	29%	Please comment
CERTH/HIT		1.66								1.66	8.74	1.66	10.4	11.76	-12%	Deviation will be compensated next period.
INESC-ID		1.2	2.7			0.1				4	15.47	4	19.47	19.78	-2%	Deviation will be compensated next period.
UNIS						1				1	20	1	21	19.98	5%	Please comment
UNIFI-MICC							4.95			4.95	27.5	4.95	32.45	24.68	31%	Development of manual annotator.
CVC				1		2			1	4	27	4	31	16.33	90%	Integration of code from different partners.
BandG							0.2	1.5		1.7	8	1.7	9.7	6.98	39%	This difference has been caused in earlier reporting.
FRD							0.1		0.2	0.3	5.35	0.3	5.65	5.52	2%	Please comment
Total	1.69	2.86	2.7	7.35	1.07	5.81	5.25	1.5	1.2	29.43	151.32	29.43	180.75	145.4	24%	

<i>Period 9, month 17 and 18 June and July 2008</i>																
Part. Short	WP1	WP2	WP3	WP4	WP5	WP6	WP7	WP8	WP9	Total	Old total	This periods to	New actual total	Planned Total	Deviation	Comment on person months
UvA managem.	0.98									0.98	12.45	0.98	13.43	11.7	15%	Please comment
UvA RTD		1		3.73	1.27	3.09				9.09	38.63	9.09	47.72	37.25	28%	Please comment
CERTH/HIT		0.64								0.64	10.4	0.64	11.04	14.04	-21%	Please comment
INESC-ID		1.2	3.1			0.1				4.4	19.47	4.4	23.87	24.28	-2%	Please comment
UNIS				0.3	0.7					1	21	1	22	24.48	-10%	Please comment
UNIFI-MICC								3		3	32.45	3	35.45	30.6	16%	Development of manual annotator not initially planned in DoW.
CVC										5		5	36		75%	Due to the intrinsic difficulty of integrating software from different partners, we had spent a lot of effort to make tests to the software. Also, a workshop is being organized, as an outcome of WP9
BandG				1		2		1		31				20.57		annotations
FRD							0.4		0.1	0.5	5.65	0.5	6.15	5.71	8%	Please comment
Total	0.98	2.84	3.1	5.03	1.97	5.19	4.4	1.87	1.1	26.48	180.75	26.48	207.23	176.82	17%	

