ITI-CERTH participation to TRECVID 2011

Anastasia Moumtzidou[1], Panagiotis Sidiropoulos[1], Stefanos Vrochidis[1,2], Nikolaos Gkalelis[1], Spiros Nikolopoulos[1,2], Vasileios Mezaris[1], Ioannis Kompatsiaris[1], Ioannis Patras[2]

[1] Informatics and Telematics Institute/Centre for Research and Technology Hellas, 1st Km. Thermi-Panorama Road, P.O. Box 60361,57001 Thermi-Thessaloniki, Greece
[2] Queen Mary, University of London, Mile End Road, London, UK
{moumtzid, psid, stefanos, gkalelis, nikolopo, bmezaris, ikom}@iti.gr, i.patras@eecs.qmul.ac.uk

Abstract

This paper provides an overview of the tasks submitted to TRECVID 2011 by ITI-CERTH. ITI-CERTH participated in the Known-item search (KIS) as well as in the Semantic Indexing (SIN) and the Event Detection in Internet Multimedia (MED) tasks. In the SIN task, techniques are developed, which combine motion information with existing well-performing descriptors such as SURF, Random Forests and Bag-of-Words for shot representation. In the MED task, the trained concept detectors of the SIN task are used to represent video sources with model vector sequences, then a dimensionality reduction method is used to derive a discriminant subspace for recognizing events, and, finally, SVM-based event classifiers are used to detect the underlying video events. The KIS search task is performed by employing VERGE, which is an interactive retrieval application combining retrieval functionalities in various modalities and exploiting implicit user feedback.

1 Introduction

This paper describes the recent work of ITI-CERTH [1] in the domain of video analysis and retrieval. Being one of the major evaluation activities in the area, TRECVID [1] has always been a target initiative for ITI-CERTH. In the past, ITI-CERTH participated in the search task under the research network COST292 (TRECVID 2006, 2007 and 2008) and in the semantic indexing (SIN) task (which is the similar to the old high-level feature extraction task) under MESH integrated project [2] (TRECVID 2008), K-SPACE project [3] (TRECVID 2007 and 2008). In 2009 and 2010 ITI-CERTH has participated as stand alone organization in the HLFE and Search tasks ([4]) and in the KIS, INS, SIN and MED tasks ([5]) of TRECVID correspondingly. Based on the acquired experience from previous submissions to TRECVID, our aim is to evaluate our algorithms and systems in order to improve and enhance them. This year, ITI-CERTH participated in three tasks: known-item search, semantic indexing and the event detection in internet multimedia tasks. In the following sections we will present in detail the applied algorithms and the evaluation for the runs we performed in the aforementioned tasks.

[1] Informatics and Telematics Institute - Centre for Research & Technology Hellas

2 Semantic Indexing

2.1 Objective of the submission

Since 2009, ITI-CERTH is working on techniques for video high-level feature extraction that treat video as video, instead of processing isolated key-frames only (e.g. [6]). The motion information of the shot, particularly local (object) motion, is vital when considering action-related concepts. Such concepts are also present in TRECVID 2011 SIN task (e.g. "Swimming", "Walking", "Car racing"). In TRECVID 2011, ITI-CERTH examines how video tomographs, which are 2-dimensional slices with one dimension in time and one dimension in space, can be used to represent the video shot content for video concept detection purposes. Two different tomograph variants were used, depending on the considered spatial dimension, namely Horizontal and Vertical tomographs. These were employed similarly to visual key-frames in distinct concept detector modules. The detector outcomes are linearly combined in order to extract the final video concept results.

Concept detector modules were built following the Bag-of-Words (BoW) scheme, using Random Forests implementation in order to reduce the associated computational time without compromising the concept detection performance. Finally, a post-processing scheme, based on the provided ontology was examined. Four full runs, denoted "ITI-CERTH-Run 1" to "ITI-CERTH-Run 4", were submitted as part of this investigation.

2.2 Description of runs

Four SIN runs were submitted in order to evaluate how the use of video tomographs [7] can enhance concept detection rate. All 4 runs were based on generating one or more Bag-of-Words models of SURF descriptors that capture 2D appearance (i.e. intensity distribution in a local neighborhood). Motion pattern is captured with the use of video tomographs, which depend on both temporal and spatial content of each shot. SURF descriptors were extracted from key-frames and tomographs following the dense sampling 64-dimensional SURF descriptor scheme introduced in [8], utilizing the software implementation of [9].

In all cases where a BoW model was defined, the number of words was set to 1024. As proposed in [8], a pyramidal 3x1 decomposition scheme was used for every key-frame, thus generating 3 different random trees. In addition a random tree using the entire image was built. Thus, a concatenated key-frame description vector of dimension 4096 was created. For the tomograph BoWs a one-level temporal pyramidal scheme was used (Fig. 1). Each shot horizontal or vertical tomograph was split into 3 equal time slots, and a BoW model was created for each one of them. Furthermore, one horizontal and one vertical BoW associated with the entire shot was created. As a result, two concatenated description vector of dimension 4096 (one for horizontal and one for vertical tomographs) were extracted for each shot.

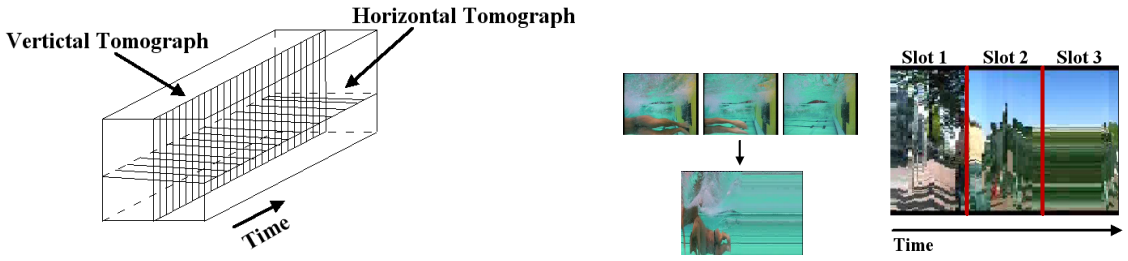

Figure 1: (a) Horizontal and vertical tomographs in a video volume (b1) Three key-frames of a shot and the resulting vertical tomograph (b2) A video shot tomograph and its decomposition into time slots.

All random forests were trained by selecting 1 million SURF vectors from 10 thousand key-frame images (or tomographs, in the case of tomograph modules) and using the training strategy that was analytically described in [8].

For the detectors, a common method was selected for all runs to provide comparable results between them. In particular, a set of SVM classifiers was trained using the different feature vectors each time. In all cases, a subset of the negative samples in the training set was selected by a random process. In order to augment the dataset with more positive samples, we extracted 9 visual key-frames, 3 horizontal tomographs and 3 vertical tomographs per shot in the training dataset. Like the 2010 competition, we used a diverse proportion of positive and negative samples for training the concept detectors. Specifically, in order to maintain computational costs at a manageable level, we set a maximum of 20000 training samples per concept. A variable proportion of positive/negative samples was used to reach the 20000 samples limit for as many concepts as possible; this proportion ranged from 1:5 to 1:1.

We have selected to implement linear kernel SVMs for two main reasons. Firstly, the size of the dataset and the number of employed concepts raised the computational cost of the unsupervised optimization procedure that is part of the LIBSVM tool [10] in prohibitive levels. Secondly, as it stated in [10], when the number of dimensions is comparable to the number of vectors to be trained (in our case the vector dimension was 4096 and the number of vectors at most 20000) the use of a more complex than linear kernel is not expected to significantly improve the SVM classifier performance.

The output of the classification for a shot, regardless of the employed input feature vector, is a value in the range [0, 1], which denotes the Degree of Confidence (DoC) with which the shot is related to the corresponding high-level feature. The outputs of key-frame as well as horizontal and vertical tomograph confidence scores were linearly combined using fixed weights that were manually selected. The final results per high-level feature were sorted by DoC in descending order and the first 2000 shots were submitted to NIST.

In one of the submitted runs, the use of the provided ontology was also tested. The adopted methodology relies on the assumption that a detector can be more easily tuned for more specific classes than for more general ones. Consequently, for those concepts that were implied by more than one more specific concepts (e.g. the concept "animal", which is implied by concepts "dog", "cat" etc.) the detection results were filtered by multiplying the estimated degree of confidence with the maximum degree of confidence of the concepts that imply it (Eq. 1).

$$FDoC_i = DoC_i * max(DoC_{i_1}, DoC_{i_2}, ...DoC_{i_n}) \quad (1)$$

where $FDoC_i$ is the final degree of confidence for concept i, which is implied by concepts $i_1, i_2, ...i_n$. The rationale behind this strategy is that, when a general concept is detected then one specific detector is also expected to return a high confidence value (e.g. if a cat is present in a shot, then besides the general concept "animal", the specific concept "cat" is also expected to return a high score).

The 4 submitted runs were:

- ITI-CERTH-Run 1: "Visual, Horizontal Tomographs and Vertical Tomographs, ontology included". This is a run combining the visual-based BoW with both the horizontal and vertical tomographs. Three visual key-frames, and one horizontal and one vertical tomograph were used to represent each shot. SURF features were extracted and a Bag-of-Words of each key-frame (or tomograph) was created using random forest binning. Visual, horizontal and vertical confidence results were linearly combined by averaging the confidence scores of all representative images (as a result, the output of the visual module accounts for 60% of the final confidence score, while each of the horizontal and vertical modules accounts for 20% of it). The results of general concepts were additionally filtered using the aforementioned ontology-based technique.

- ITI-CERTH-Run 2: "Visual, Horizontal Tomographs and Vertical Tomographs". This run is similar to run 1, the only difference being that the final ontology-based post-processing step is omitted.

- ITI-CERTH-Run 3: "Visual and Vertical Tomographs". This is a run combining the visual-based BoW with only vertical tomographs. Three visual key-frames and one vertical tomograph were used to represent each shot. SURF features were extracted and a Bag-of-Words of each key-frame (tomograph) was created using random forest binning. Visual and vertical confidence results were linearly combined by averaging the confidence scores of all representative images (as a result, the output of the visual module accounts for 75% of the final confidence score, while the vertical module accounts for 25% of it).

- ITI-CERTH-Run 4: "Visual". This is a baseline run using only three visual key-frames and averaging operation of their confidence scores. SURF descriptors and random forests were employed, as in the previous run

2.3 Results

The runs described above were submitted for the 2011 TRECVID SIN competition. The evaluation results of the aforementioned runs are given in terms of the Mean Extended Inferred Average Precision (MXinfAP) both per run and per high level feature. Table 1 summarizes the results for each run presenting the Mean Extended Inferred Average Precision of all runs.

Table 1: Mean Extended Inferred Average Precision for all high level features and runs.

	ITI-CERTH 1	**ITI-CERTH 2**	**ITI-CERTH 3**	**ITI-CERTH 4**
MxinfAP	**0.042**	**0.039**	**0.041**	**0.036**
MxinfAP Light	**0.025**	**0.026**	**0.026**	**0.023**

The "Visual" run (ITI-CERTH run 4) was the baseline run of the submission. It combines SURF-based bag-of-words with random forests and spatial pyramidal decomposition to establish a baseline performance run. The "Visual and Vertical Tomographs" run (ITI-CERTH run 3) is used to assess the use of the video tomographs for concept detection. This technique did show some performance gain, improving by 15% the overall performance of the baseline run.

ITI-CERTH run 2 incorporates both horizontal and vertical tomographs. It performed better than the baseline run but worse than run 3, which employs only vertical tomographs. The inability of horizontal tomographs to represent the motion content of a shot can be explained by the fact that, in contrary to vertical tomographs that can capture the extensively used horizontal movement of a camera or of an object, horizontal tomographs capture more or less random vertical movements. However, when the results of run 2 were post-processed using the provided ontology (run 1), the performance increased by another 10%, making run 1 the best scoring run of our experiments.

It should be noted that after submitting the runs we have discovered a major bug in the tomograph extraction process, which is expected to adversely affect the performance. We are currently rebuilding the concept detectors and we hope to be able to report the correct tomograph influence to video concept detection by the time that the TRECVID conference takes place.

3 Event Detection in Internet Multimedia

3.1 Objective of the submission

The recognition of high level events in video sequences is a challenging task that is usually realized with the help of computationally demanding algorithms. For applications that require low-latency response times, such as multimedia management applications, the training and especially the testing time of pattern detection algorithms is a very critical quality factor. The objective of our participation in TRECVID MED 2011 is to evaluate the effectiveness of our computationally efficient event detection algorithm. This algorithm uses in our experiments only limited video information, i.e., one keyframe per shot and only static visual feature information.

3.2 Description of submitted run

In this section, we first give an overview of our event detection method, then describe the TRECVID MED 2011 dataset, and finally provide the implementation details of our submission.

3.2.1 Overview of the event detection method

The main parts of our event detection method are briefly described in the following:

Video representation: At the video preprocessing stage, the shot segmentation algorithm described in [11] is applied to each video for segmenting it to shots, and then F trained concept detectors are used for associating each shot with a model vector [12, 13]. More specifically, given a set $\mathcal{G}$ of F trained concept detectors, $\mathcal{G} = \{(d_\kappa(), h_\kappa),\ \kappa = 1, \ldots, F\}$, where $d_\kappa()$ is the κ-th concept detector functional and h_κ is the respective concept label, the p-th video in the database is expressed as $\mathbf{X}_p = [\mathbf{x}_{p,1}, \ldots, \mathbf{x}_{p,l_p}], \mathbf{X}_p \in \mathbb{R}^{F \times l_p}$, where $\mathbf{x}_{p,q} = [x_{p,q,1}, \ldots, x_{p,q,K}]^T,\ \mathbf{x}_{p,q} \in \mathbb{R}^F$ is the model vector associated with the q-th shot of the p-th video.

Discriminant analysis : A large number of concepts may not be relevant with the target events. To this end, a discriminant analysis (DA) technique can be used to implicitly extract the concepts that are relevant to the underlying events. For this, we apply a variant of the mixture subclass discriminant analysis (MSDA) [14] to derive a lower dimensional representation of the videos. In more detail, using the set of the training model vectors $\{(\mathbf{x}_{p,q}, y_p), p = 1, \ldots, L,\ q = 1, \ldots, l_p\}$, where y_p is the event label, a transformation matrix $\mathbf{W} \in \mathbb{R}^{F \times D}$, $D << F$, is computed so that a model vector $\mathbf{x}_{p,q}$ can be represented with $\mathbf{z}_{p,q} \in \mathbb{R}^D$ in the discriminant subspace, i.e., $\mathbf{z}_{p,q} = \mathbf{W}^T \mathbf{x}_{p,q}$.

Event recognition : The set of the training model vectors in the discriminant subspace, $\{(\mathbf{z}_{p,q}, y_p)$, $p = 1, \ldots, L,\ q = 1, \ldots, l_p\}$ is used to train one support vector machine (SVM) for each event. For the training, the one-against-all method is applied, that is, the i-th SVM, s_i, associated with the i-th event is trained considering all model vectors that belong to the i-th event as positive samples and the rest of the model vectors as negative samples. During the training procedure, along with the SVM parameters, a threshold value θ_i is also identified with respect to the i-th SVM-based event detector, which is used to transform the DoC in the output of the SVMs to a hard decision regarding the presence of an event or not in the video shot. At the evaluation stage, the j-th test video is first segmented to its constituent shots, the concept detectors are used to represent the video with a sequence of model vectors, and the MSDA projection matrix is applied to represent the video in the discriminant subspace as a sequence of projected model vectors $\mathbf{z}_{j,1}, \ldots, \mathbf{z}_{j,l_t}$. For the detection of the i-th event in the test video, the i-th event detector is then applied to produce a set of DoCs, $\delta^i_{t,1}, \ldots, \delta^i_{t,l_t}$, for each video shot, and then the following rule is applied for deciding whether the event is depicted in the video:

$$\text{median}\{\delta^i_{t,1}, \ldots, \delta^i_{t,l_t}\} > \theta_i \tag{2}$$

That is, the i-th event is detected if the median of the DoCs is larger than the threshold θ_i related to the i-th event.

3.2.2 Dataset description

The TRECVID MED 2011 evaluation track provides a new Internet video clip collection of more than 1570 hrs of clips that contains 15 events and a very large number of clips belonging to uninteresting events. The events are separated to training events, which are designated for training purposes, and to testing events, which are used for evaluating the performance of the event detection methods (Table 2). The overall dataset is divided to three main data sets: a)the EVENTS set that contains the 15 event kits, b) the transparent development collection (DEVT) that contains clips for facilitating the training procedure, and, c) the opaque development set (DEVO). The two former sets (EVENTS, DEVT) are designated for training the event detection algorithms, while the latter (DEVO) is used for the blind evaluation of the algorithms. The ground truth annotation tags used for declaring the relation of a video clip to a target event are "positive", which denotes that the clip contains at least one instance of the event, "near miss", to denote that the clip is closely related to the event but it lacks critical evidence for a human to declare that the event occurred, and, "related" to declare that the clip contains one or more elements of the event but does not meet the requirements to be a positive event instance. In case that the clip is not related with any of the target events the label "NULL" is used. Besides the clearly uninteresting videos, for training purposes we treated the clips that are annotated as "near miss" or "related", regarding a target event also as negative instances of the event, i.e., as events that belong to the uninteresting events category. Using those annotation conventions, the distribution of the clips along the testing events in all datasets are shown in Table 3.

Table 2: TRECVID MED 2011 testing and training events.

Training events	*Testing events*
E001: Attempting a board trick	E006: Birthday party
E002: Feeding an animal	E007: Changing a vehicle tire
E003: Landing a fish	E008: Flash mob gathering
E004: Wedding ceremony	E009: Getting a vehicle unstuck
E005: Working on a woodworking project	E010: Grooming an animal
	E011: Making a sandwich
	E012: Parade
	E013: Parkour
	E014: Repairing an appliance
	E015: Working on a sewing project

Table 3: TRECVID MED 2011 video collection.

EVEN. ID	*E001*	*E002*	*E003*	*E004*	*E005*	*E006*	*E007*	*E008*
EVENTS	160	161	119	123	141	172	110	173
DEVT	137	125	93	90	99	10	2	1
DEVO	–	–	–	–	–	186	111	132
	E009	*E010*	*E011*	*E012*	*E013*	*E014*	*E015*	*Other*
	128	137	124	136	111	121	120	356
	–	–	1	21	2	9	9	10122
	95	87	140	231	104	78	81	30576

Table 4: Detection thresholds.

Event ID	*E006*	*E007*	*E008*	*E009*	*E010*	*E011*	*E012*	*E013*	*E014*	*E015*
Thresh.	0.7200	0.5300	0.6500	0.5670	0.5500	0.6600	0.6400	0.6100	0.6700	0.5000

3.2.3 Experimental setup

We first utilize the automatic segmentation algorithm described in [11] for the temporal decomposition of the videos in the EVENTS and DEVT sets to video shots. We then select one keyframe per shot to represent each video with a sequence of shot keyframes and apply the model vector-based procedure to represent each shot keyframe with a 346-dimensional model vector. This is done by firstly extracting keypoints from each keyframe and using them to form 64-dimensional SURF descriptor vectors [15] and then following the concept detection method described in Section 2 (SIN task, run 4). The output of each concept detector is a number in the range $[0, 1]$ expressing the degree of confidence (DoC) that the concept is present in the keyframe. The values of all the detectors are concatenated in a vector, to yield the model vector representing the respective shot. Consequently, the whole set of training model vectors is used for optimizing the parameters of the MSDA algorithm (e.g., dimensionality of output vectors) as well as the parameters of the kernel SVMs that are used for event detection, and for identifying the event specific thresholds (Table 4). The overall optimization procedure was guided by the Normalized Detection Cost (NDC), i.e., NDC was the quantity to be minimized. During the testing stage, the same procedure is followed to represent each video in the DEVO collection with the respective sequence of model vectors. The model vector sequences are then projected in the discriminant subspace using the MSDA projection matrix, and the SVM-based event detectors along with the median rule and the event specific thresholds are applied for the detection of the target events.

Table 5: Evaluation results.

Event ID	*TP*	*FA*	P_{FA}	P_{MS}	*NDC*
E006	1	421	0.0133	0.9946	1.1608
E007	13	1706	0.0538	0.8829	1.5547
E008	17	1120	0.0353	0.8712	1.3126
E009	25	1691	0.0533	0.7368	1.4024
E010	15	1298	0.0409	0.8276	1.3384
E011	3	802	0.0253	0.9786	1.2947
E012	25	797	0.0252	0.8918	1.2068
E013	6	2517	0.0794	0.9423	1.9333
E014	2	101	0.0032	0.9744	1.0141
E015	11	1856	0.0585	0.8642	1.5944

3.3 Results

The evaluation results of the run described above are given in Table 5, in terms of true positives (TP), False Alarms (FA), false alarms rate (P_{FA}), missed detections rate (P_{MS}) and actual Normalized Detection Cost (NDC). Figure 2 depicts the evaluation results of all submissions in terms of the average actual NDC along all ten evaluation events, while in Figure 3 the same information is provided but now only for the submissions that exploit exclusively visual information. We observe that we succeed rather average performance compared to the other submissions. This is expected as we use only static visual information (SURF) and exploit only one keyframe per shot, in contrast to the majority of the other participants that utilize many keyframes per shot, and exploit several sources of information, such as color features (e.g., OpponentSift, color SIFT), motion features (e.g., STIP, HOF), audio features (e.g., MFCC, long-term audio texture), and other. Therefore, we can conclude that although limited video information is exploited (sparsely sampled video sequences and static visual features) our method still achieves average detection performance.

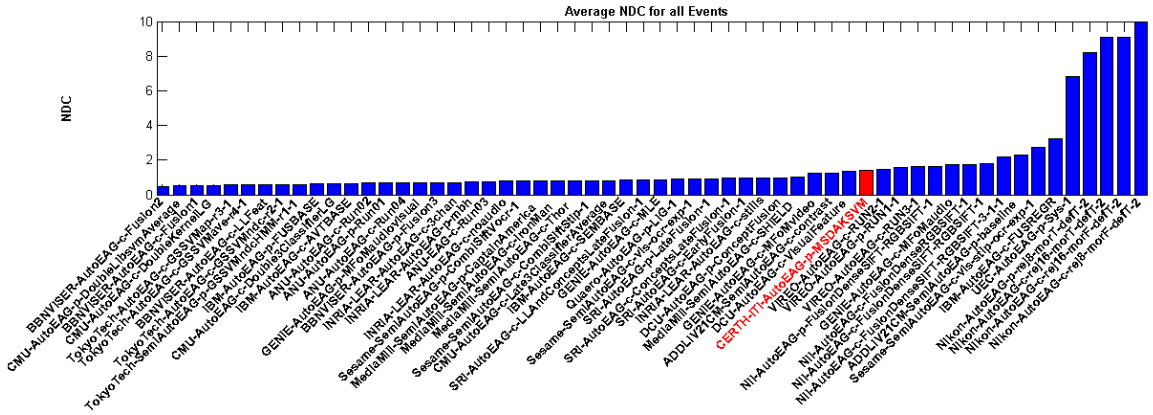

Figure 2: Average actual NDC for all submissions.

In terms of computational complexity, excluding any processes that are related to other TRECVID tasks (e.g., extraction of concept detection values using the SIN task method), the application of the MSDA method combined with the SVM-based event detection process to the precomputed model vector sequences is executed in real time. For instance, the computational times for applying these techniques in the overall DEVO collection for the detection of the 10 evaluation events requires a few minutes as shown in Table 6.

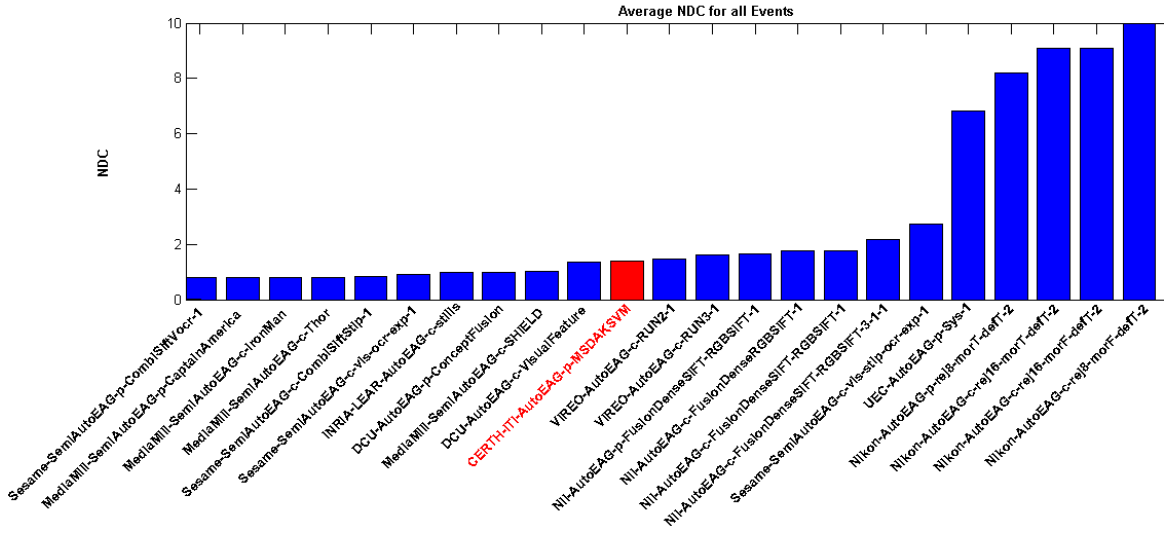

Figure 3: Average actual NDC for all submissions that use only visual information.

Table 6: Event agent execution times for MSDA and SVM-based event detector.

Event ID	*E006*	*E007*	*E008*	*E009*	*E010*	*E011*	*E012*	*E013*	*E014*	*E015*
Time (mins)	5.8	6.277	13.605	4.197	4.858	10.492	8.532	7.721	9.08592	7.342

4 Known Item Interactive Search

4.1 Objective of the submission

ITI-CERTH's participation in the TRECVID 2011 known-item (KIS) task aimed at studying and drawing conclusions regarding the effectiveness of a set of retrieval modules, which are integrated in an interactive video search engine. Within the context of this effort, several runs were submitted, each combining existing modules in a different way, for evaluation purposes.

Before we proceed to the system description we will provide a brief description of KIS task. As it is defined by TRECVID guidelines, the KIS task represents the situation, in which the user is searching for one specific video contained in a collection. It is assumed that the user already knows the content of the video (i.e. he/she has watched it in the past). In this context, a detailed textual description is provided to the searchers accompanied with indicative keywords.

4.2 System Overview

The system employed for the Known-Item search task was VERGE[2], which is an interactive retrieval application that combines basic retrieval functionalities in various modalities, accessible through a friendly Graphical User Interface (GUI), as shown in Figure 4. The following basic modules are integrated in the developed search application:

- Implicit Feedback Capturing Module;
- Visual Similarity Search Module;
- Transcription Search Module;
- Metadata Processing and Retrieval Module;
- Video Indexing using Aspect Models and the Semantic Relatedness of Metadata;

- High Level Concept Retrieval and Fusion Module;
- High Level Concept and Text Fusion Module;

The search system is built on open source web technologies, more specifically Apache server, PHP, JavaScript, mySQL database, Strawberry Perl and the Indri Search Engine that is part of the Lemur Toolkit [16].

Besides the basic retrieval modules, VERGE integrates a set of complementary functionalities, which aim at improving retrieved results. To begin with, the system supports basic temporal queries such as the shot-segmented view of each video, as well as a shot preview by rolling three different keyframes. The selected shots by a user could be stored in a storage structure that mimics the functionality of the shopping cart found in electronic commerce sites. Finally, a history bin is supported, in which all the user actions are recorded. A detailed description of each of the aforementioned modules is presented in the following sections.

4.2.1 Implicit Feedback Capturing Module

A new feature added in the current version of VERGE is the recording of human-machine interaction with a view to exploiting the implicit user feedback. More specifically, the idea is to identify the shots that are of interest to the user in order to tune the retrieval modules accordingly. In this case we have considered as the main implicit interest indicator the time duration that a user is hovering over a shot to preview it. Given the fact that the user is searching to find a specific video, it is highly possible that when he/she previews a shot, the latter has common characteristics with the desirable video. To this end we record all the shots previewed by the user during the same search session, in which the user is searching for the same topic. The approach we followed is based on the assumption that there are topics for which specific visual concepts are important (or perform better than others) and cases that ASR or metadata are more important compared to visual concepts or the opposite. Therefore we suggest that the implicit information could be used in order to train weights between different modalities or between instances of the same modality and generate a more intelligent fusion as proposed in [17]. In this context we have implemented a fusion model to combine results for different concepts, as well as between textual information (metadata or ASR) and visual concepts. The fusion techniques will be described in detail in sections 4.2.6 and 4.2.7.

4.2.2 Visual Similarity Search Module

The visual similarity search module performs image content-based retrieval with a view to retrieving visually similar results. Following the visual similarity module implementation in [5], we have chosen two MPEG-7 schemes: the first one relies on color and texture (i.e., ColorLayout and EdgeHistogram were concatenated), while the second scheme relies solely on color (i.e., ColorLayout and ColorStructure).

4.2.3 Transcription Search Module

The textual query module exploits the shot audio information. To begin with, Automatic Speech Recognition (ASR) is applied on test video data. In this implementation, the ASR is provided by [18]. The textual information generated is used to create a full-text index utilizing Lemur [16], a toolkit designed to facilitate research in language modelling.

4.2.4 Metadata Processing and Retrieval Module

This module exploits the metadata information that is associated with the videos. More specifically, along with every video of the collection, an XML file is provided that contains a short metadata description relevant to the content of the video. The fist step of the metadata processing involves the parsing of the XML files and particularly the extraction of the content located inside the following tags: title, subject, keywords and description. The next step deals with the processing of the acquired content and includes punctuation and stop words removal. Finally, the processed content was indexed

[2]VERGE: http://mklab.iti.gr/verge

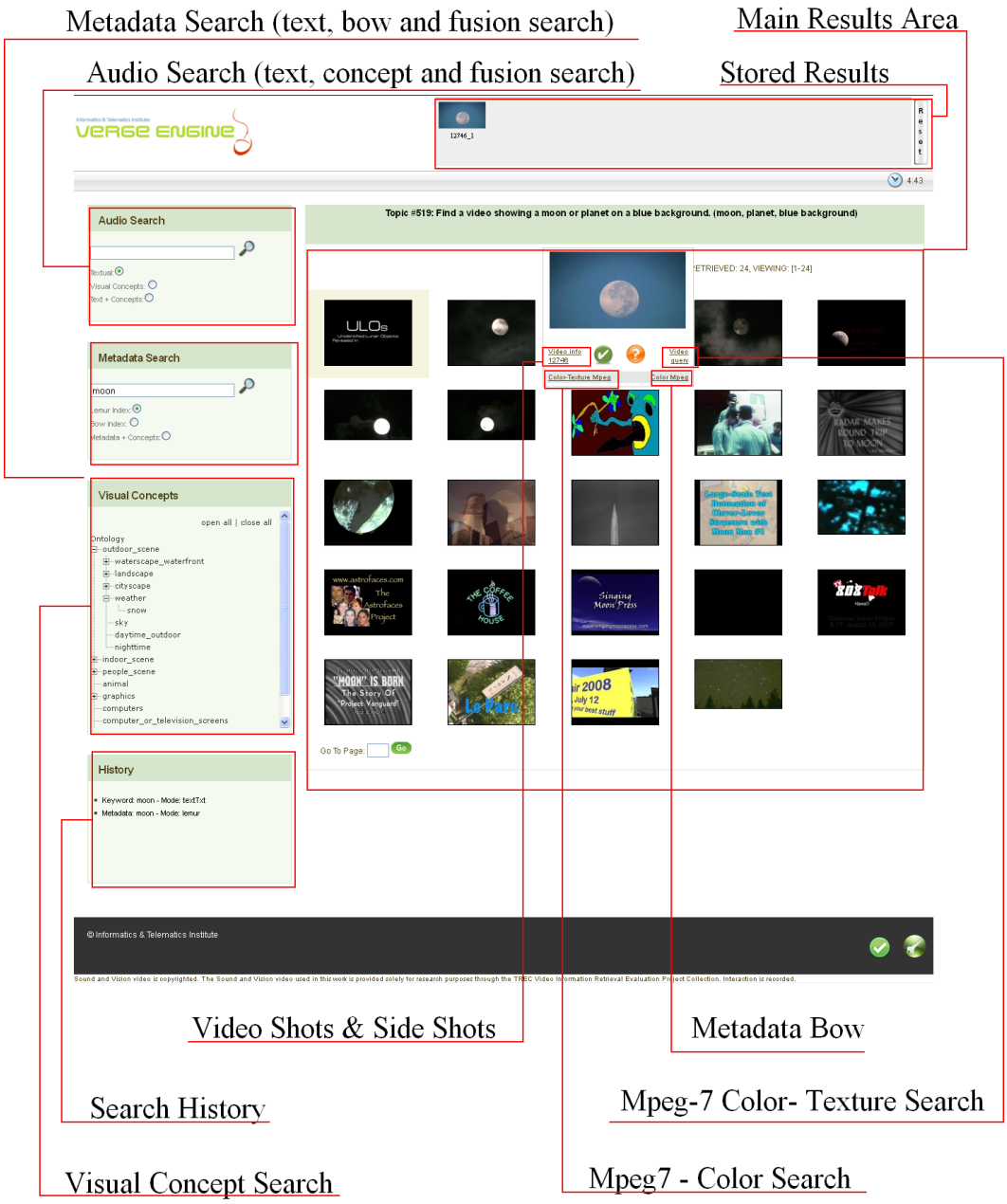

Figure 4: User interface of the interactive search platform and focus on the high level visual concepts.

with the Lemur toolkit that enables fast retrieval as well easy formulation of complicated queries in the same way described in section 4.2.3.

4.2.5 Video indexing using aspect models and the semantic relatedness of metadata

For implementing the "Video Query" functionality we have employed a bag-of-words (BoW) representation of video metadata. More specifically, in order to express each video as a bag-of-words we initially pre-processed the full set of metadata for removing stop words and words that are not recognized by WordNet [19]. Then, by selecting the 1000 most frequent words to define a Codebook of representative words, we have expressed each video as an occurrence count histogram of the representative words in its metadata. Subsequently, in order to enhance the semantic information enclosed by the bag-of-words representation, we have used a WordNet-based similarity metric [20] to measure the semantic relatedness of every word in the Codebook with all other members of the Codebook. In this way, we have managed to generate a matrix of semantic similarities, that was used to multiply the bag-of-words representations of all videos. Finally, probabilistic Latent Semantic Analysis [21] was applied on the semantically enhanced video representations to discover their hidden relations. The result of pLSA was to express each video as a mixture of 25 latent topics, suitable for performing indexing and retrieval on the full video collection.

For indexing new video descriptions, such as the as the ones provided by the user in the "Transcription Search Module", the pLSA theory proposes to repeat the Expectation Maximization (EM) steps [22] that have been used during the training phase, but without updating the values of the word-topic probability distribution matrix. However, due to some technical constraints of our implementation environment we have adopted a more simplistic approach. More specifically, we have transformed the user-provided video description into the space of latent topics by simply multiplying the semantically enhanced BoW representation of description with the word-topic probability distribution matrix. Although convenient for our implementation environment, our experimental findings has proven this solution to be sub-optimal from the perspective of efficiency.

4.2.6 High Level Visual Concept Retrieval and Fusion Module

This module facilitates search by indexing the video shots based on high level visual concept information such as water, aircraft, landscape, crowd. Specifically, we have incorporated into the system all the 346 concepts studied in the TRECVID 2011 SIN task using the techniques and the algorithms described in detail in section 2. It should be noted that in order to expand the initial set of concepts, we inserted manually synonyms that could describe the initial entries equally well (e.g. as synonyms of the concept "demonstration" were considered "protest" and "riot". In order to combine the results provided by several concepts we applied late fusion by employing the attention fusion model suggested in [17]. The following formula was used to calculate the similarity score among the query concept (q) and a shot/document (D).

$$R(q,D) = \frac{R_{avg} + \frac{1}{2(n-1)+n\gamma}\sum_i |n\omega_i R_i(q,D) - R_{avg}|}{W} \tag{3}$$

where

$$R_{avg} = \sum_i \omega_i R_i(q,D) \tag{4}$$

and

$$W = 1 + \frac{1}{2(n-1)+n\gamma}\sum_i |1 - n\omega_i| \tag{5}$$

In the previous formulas, γ is a predefined constant and it is fixed to 0.2, n is the number of modalities (i.e. in the case of concept fusion it is set to the different number of concepts) and w_i is the weight of each modality. Moreover, R_i reflects the relevance for each modality and for each shot.

In the case that there isn't any feedback from the user, we use equal normalized weights for each concept. When implicit user feedback is available, the weight for each concept is obtained from the following formula:

$$w = \sum_i t_i c_i \tag{6}$$

where c_i stands for the normalized DoC of the specific shot i and concept and t_i for the normalized attention weight of the specific shot i. Afterwards, the weights w of each concept are normalized by dividing their value with their sum from all concepts. The weights are constantly updated as the user is viewing more shots during the search session while the time duration threshold for which a shot was considered "previewed" was set to 700 milliseconds.

4.2.7 High Level Concepts and Text Fusion Module

This module combines the textual, either audio or metadata information, with the high level visual concepts of the aforementioned modules. Two cases of fusion where considered: i) visual concepts and text from ASR and ii) visual concepts and metadata. In the first case the fusion was realized at shot level. During the first minutes of interaction we applied the attention fusion model using again the formula of 3. After a reasonable number of examples has been identified (in this case we have set the threshold to 7) we applied a linear SVM regression model [23] using as features the normalized results from the different modalities and normalized relevance scores proportional to the preview time. We employed the same fusion methodology for the second case. However, the metadata involved refer to the whole video and not to specific shots. Therefore we have realized a fusion at video level by generating concept scores for each video. Based on the assumption that the important information for a video is whether a concept exists or not (and not how many times it appears) we simply assigned the greater confidence value between the shots of one video for a certain concepts.

4.3 Known-Item Search Task Results

The system developed for the known-item search task includes all the aforementioned modules apart from the segmentation module. We submitted four runs to the Known-Item Search task. These runs employed different combinations of the existing modules as described below:

Table 7: Modules incorporated in each run.

Modules	Run IDs) I_A_YES_ITI-CERTH_x			
	x=1	x=2	x=3	x=4
ASR Lemur text	no	yes	yes	yes
ASR fusion	yes	no	no	no
Metadata Lemur text	no	yes	yes	no
Metadata BoW text	no	no	yes	yes
Metadata fusion	yes	no	no	no
High Level Visual concepts	yes	yes	no	yes

The complementary functionalities were available in all runs, while the time duration for each run was considered to be five minutes. The number of topics and the mean inverted rank for each run are illustrated in Table 8.

Table 8: Evaluation of search task results.

Run IDs	Mean Inverted Rank	0.560
I_A_YES_ITI-CERTH_1	0.560	14/25
I_A_YES_ITI-CERTH_2	0.560	14/25
I_A_YES_ITI-CERTH_3	0.560	14/25
I_A_YES_ITI-CERTH_4	0.320	8/25

By comparing the values of Table 8, we can draw conclusions regarding the effectiveness of each of the aforementioned modules. The first 3 runs achieved the same score despite the different search options provided. On the other hand the 4th run achieved a lower score due to the fact that the full text metadata search option was not available and the simplistic approach followed for metadata search in this case (described at section 4.2.5) did not perform very well. The runs 2 and 3 achieved the same score showing that the visual concepts didn't help the users in retrieving better results. The same conclusion can be made when we compare the latter with run 1 as despite the text and concept fusion the results were not improved. Compared to the other systems participated in interactive Known Item Search, three of our runs achieved the best score reported in this year's KIS task, while only one run from another system achieved the same score.

5 Conclusions

In this paper we reported the ITI-CERTH framework for the TRECVID 2011 evaluation. ITI-CERTH participated in the SIN, KIS and MED tasks in order to evaluate existing techniques and algorithms.

Regarding the TRECVID 2010 SIN task, a large number of new high level features has been introduced, with some of them following a significant motion pattern. In order to take advantage of the motion activity in each shot we have extracted 2-dimensional slices, named tomographs, with one dimension in space and one in time. The use of these tomographs, as well as the provided ontology resulted to an improvement of 16.7% over the baseline approach.

As far as KIS task is concerned, the results reported were satisfactory and specific conclusions were drawn. First, the full text ASR and metadata search were the most effective retrieval modules, while visual concept retrieval didn't provide an added value. Fusion of different modalities could be promising, however we cannot draw safe conclusions due to the limited search session time and the low performance (due to the aforementioned bug) of visual concepts. Regarding the BoW-based metadata retrieval module, it didn't have a high impact to the results due to the simplistic implementation attempted.

Finally, as far as the TRECVID 2011 MED task is concerned a "model vector-based approach", combined with a dimensionality reduction method and a set of SVM-based event classifiers has been evaluated. The proposed approach provided an average detection performance, exploiting however only basic visual features and a sparse video representation. This event detection approach is advantageous in terms of computational complexity as discussed above.

6 Acknowledgements

This work was partially supported by the projects GLOCAL (FP7-248984) and PESCaDO (FP7-248594), both funded by the European Commission.

References

[1] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[2] MESH, Multimedia sEmantic Syndication for enHanced news services. `http://www.mesh-ip.eu/?Page=project`.

[3] K-Space, Knowledge Space of Semantic Inference for Automatic Annotation and Retrieval of Multimedia Content. `http://kspace.qmul.net:8080/kspace/index.jsp`.

[4] A. Moumtzidou, A. Dimou, P. King, and S. Vrochidis et al. ITI-CERTH participation to TRECVID 2009 HLFE and Search. In *Proc. TRECVID 2009 Workshop*, pages 665–668. 7th TRECVID Workshop, Gaithersburg, USA, November 2009, 2009.

[5] A. Moumtzidou, A. Dimou, N. Gkalelis, and S. Vrochidis et al. ITI-CERTH participation to TRECVID 2010. In *Proc. TRECVID 2010 Workshop*. 8th TRECVID Workshop, Gaithersburg, MD, USA, November 2010, 2010.

[6] J. Molina, V. Mezaris, P. Villegas, and G. Toliasand E. Spyrou et al. Mesh participation to trecvid2008 hlfe. 6th TRECVID Workshop, Gaithersburg, USA, November 2008, 2008.

[7] Sebastian Possos and Hari Kalva. Accuracy and stability improvement of tomography video signatures. In *ICME 2010*, pages 133–137, 2010.

[8] Jasper Uijlings, Arnold Smeulders, and Remko Scha. Real-time visual concept classification. *IEEE Transactions on Multimedia*, 12(7):665–681, 2010.

[9] Ork de Rooij, Marcel Worring, , and Jack van Wijk. Mediatable: Interactive categorization of multimedia collections. *IEEE Computer Graphics and Applications*, 30(5):42–51, 2010.

[10] C. C. Chang and C. J. Lin. *LIBSVM: a library for support vector machines*, 2001.

[11] E. Tsamoura, V. Mezaris, and I. Kompatsiaris. Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. In *Proc. IEEE Int. Conf. on Image Processing, Workshop on Multimedia Information Retrieval (ICIP-MIR 2008)*, pages 45–48. San Diego, CA, USA, October 2008, 2008.

[12] V. Mezaris, P. Sidiropoulos, A. Dimou, and I. Kompatsiaris. On the use of visual soft semantics for video temporal decomposition to scenes. In *Proc. Forth IEEE Int. Conf. on Semantic Computing (ICSC 2010)*, pages 141–148, Pittsburgh, PA, USA, September 2010.

[13] J. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME '03)*, pages 445–448, Baltimore, MD, USA, July 2003.

[14] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. Mixture subclass discriminant analysis. 18(5):319–332, May 2011.

[15] H. Bay, A. Ess, T. Tuytelaars, and L. Vangool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, June 2008.

[16] The lemur toolkit. `http://www.cs.cmu.edu/ lemur`.

[17] Bo Yang, Tao Mei, Xian-Sheng Hua, Linjun Yang, Shi-Qiang Yang, and Mingjing Li. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 73–80, New York, NY, USA, 2007. ACM.

[18] Julien Despres, Petr Fousek, Jean-Luc Gauvain, Sandrine Gay, Yvan Josse, Lori Lamel, , and Abdel Messaoudi. Modeling Northern and Southern Varieties of Dutch for STT. In *Interspeech 2009*, pages 96–99, Brighton, UK, September 2009.

[19] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

[20] Siddharth Patwardhan. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Master's thesis, August 2003.

[21] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.

[22] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley and Sons, 2nd edition, 1997.

[23] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.