ITI-CERTH participation to TRECVID 2009 HLFE and Search

Anastasia Moumtzidou, Anastasios Dimou, Paul King, Stefanos Vrochidis,
Angeliki Angeletou, Vasileios Mezaris, Spiros Nikolopoulos, Ioannis Kompatsiaris,
Lambros Makris

*Informatics and Telematics Institute/Centre for Research and Technology Hellas, 1st Km.
Thermi-Panorama Road, P.O. Box 60361, 57001 Thermi-Thessaloniki, Greece*
{moumtzid, dimou, king, stefanos, angelet, bmezaris, nikolopo, ikom, lmak}@iti.gr

February 17, 2010

Abstract

This paper provides an overview of the tasks submitted to TRECVID 2009 by ITI-CERTH. ITI-CERTH participated in the high-level feature extraction task and the search task. In the high-level feature extraction task, techniques are developed that combine motion information with existing well-performing descriptors such as SIFT and Bag-of-Words for shot representation. In a separate run, the use of compressed video information to form a Bag-of-Words model for shot representation is studied. The search task is based on an interactive retrieval application combining retrieval functionalities in various modalities (i.e. textual, visual and concept search) with a user interface supporting interactive search over all queries submitted. Evaluation results on the submitted runs for this task provide interesting conclusions regarding the comparison of the involved retrieval functionalities as well as the strategies in interactive video search.

1 Introduction

This paper describes the work of ITI-CERTH [1] in the area of video analysis and retrieval. Being one of the major evaluation activities in the area, TRECVID 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 [1, 2, 3]) and in the high level feature extraction task (HLFE) under MESH integrated project [4] (TRECVID 2008 [5]) and K-SPACE project [6] (TRECVID 2007 and 2008 [7, 8]). Therefore, based on the acquired experience from previous submissions to TRECVID, we have tried to improve and enrich our algorithms and systems. This year, ITI-CERTH participated in two tasks: high-level feature extraction and search. The following sections detail the applied algorithms and their evaluations.

2 High-level feature extraction

2.1 Objective of the submission

The participation of ITI-CERTH in the TRECVID HLFE task over the past few years followed the mainstream approach of processing one or more representative keyframes for each shot in order to extract the high-level features (e.g. [5]). However, it is evident that a keyframe represents only a small portion of the information contained in a shot. The motion information of the shot in particular (which includes information on camera motion and local object motion) is ignored when considering only isolated keyframes. This observation, together with the fact that at least some of the high-level

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

features defined in the 2009 competition were action-related (e.g. “Person-playing-soccer”, “Person-riding-a-bicycle”, “Traffic intersection”) rather than static (e.g. “outdoors”), motivated ITI-CERTH to examine in TRECVID 2009 how motion information, extracted from processing the entire shot instead of selected keyframes, can be combined with well-performing techniques such as SIFT and Bag-of-Words (BoW) towards improved high-level feature extraction in video. Five runs, denoted “ITI-CERTH-Run 1” to “ITI-CERTH-Run 5”, were submitted as part of this investigation.

In addition to the above, ITI-CERTH also participated in the TRECVID HLFE task through the MESH-ITI-CERTH submission with one run. The objective of this run was to examine the potential of directly using compressed video information to form a Bag-of-Words model for representing the video shots. The motivation behind it was that, in practice, the application of the (commonplace in TRECVID) shot description method that is based on extracting SIFT descriptors from one or more keyframes of a shot requires the decoding of the (typically MPEG-2) video, the extraction of the keyframes, and the application of techniques for interest point detection and description to each keyframe. All of these processes are relatively computationally intensive, making the real-time processing of video rather problematic. In contrast, the possibility of directly using compressed domain information to form the BoW model would allow for much-faster-than-real-time processing of the videos, enabling the application of high-level feature extraction techniques to unprecedented volumes of video data.

2.2 Description of runs

Five HLFE runs were submitted in order to evaluate how motion information can enrich the traditional BoW technique, and one additional HLFE run in order to examine the potential of directly using compressed video information. All 6 runs were based on generating one or more Bag-of-Words models, each time using different features that may represent motion information in addition to the intensity distribution in a local neighbourhood that SIFT descriptors represent, or global compressed video information. In all cases where SIFT descriptors were extracted from an image, using the original 128-dimensional SIFT descriptors introduced in [9], the software implementation of [10] was utilized. Additionally, in all cases where a BoW model was defined, the number of words was set to 500. This is quite low for TRECVID standards, where BoW models of 4000 words or more are typically used; nevertheless, it allowed for the timely completion of the experiments on typical PCs and, at the same time, this choice did not negatively impact the desired comparison between the 6 different runs of the ITI-CERTH and MESH-ITI-CERTH submissions.

A common training method for all runs was selected 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. It was employed, instead of a complete set, in a 5:1 proportion to the positive samples available for each high-level feature, to facilitate the training of the classifiers in cases where the number of positive samples available for training was disproportionately low. The SVM parameters were set by an unsupervised optimization procedure that is part of the LIBSVM tool [11]. The output of the classification for a shot, regardless of the employed input feature vector, is a number in the continuous range [0, 1] expressing the Degree of Confidence (DoC) that the shot relates to the corresponding high-level feature. The results per high-level feature were sorted by DoC in descending order and the first 2000 shots were submitted to NIST.

The 6 submitted runs were:

- ITI-CERTH-Run 1: “SIFT”. This is a baseline run using the traditional Bag-of-Words methodology on SIFT features extracted from one representative keyframe of the shot.

- ITI-CERTH-Run 2: “MSIFT”. In this run, instead of using SIFT features extracted from a representative keyframe to generate the BoW, SIFT extraction was performed on all frames of a temporally sub-sampled version of each shot; tracks of similar interest points (i.e., interest points in successive frames of the temporally sub-sampled shot that had similar SIFT descriptors and locations on the image grid, taking into account the camera motion) were identified, and the element-wise average SIFT vectors of a selected sub-set of the identified interest point tracks were used to generate the BoW model.

- ITI-CERTH-Run 3: "MTD". In this run, each of the average 128-element SIFT vectors of a selected sub-set of the interest point tracks identified in the shot, as in run 2, was concatenated with a 112-element vector describing the direction of motion of the interest point track across different temporal scales and different partitionings of the $[-\frac{\pi}{2}, \frac{\pi}{2}]$ motion direction space. The resulting 240-element vectors were then used for generating the BoW model.

- ITI-CERTH-Run 4: "SMS". This run used as input to the SVMs the combination of the shot descriptions of runs 1 and 2. In particular, the BoW descriptors of each shot according to run 1 and run 2 were estimated separately, as in the aforementioned runs, and were subsequently concatenated to form a 1000-element SVM input vector (as opposed to the 500-element SVM input vectors of runs 1 to 3).

- ITI-CERTH-Run 5: "SMT". This was similar to run 4, the difference being that the BoW descriptors of each shot according to run 1 and run 3 were concatenated and served as input to the SVMs, instead of the run 1 and run 2 descriptors.

- MESH-ITI-CERTH-Run 1: "DCT". In this run, the use of compressed video information for BoW model generation was examined for the first time. The 384-element vectors of DCT coefficients of all macroblocks (MBs) of one intra-coded frame per shot were used in place of the SIFT descriptors that would have typically been extracted from a fully decoded keyframe. The 384 elements of each vector corresponded to the 256 DCT coefficients describing the Y component of the MB, and 64 DCT coefficients for each one of the Cb and Cr components. Subsequently, the usual BoW methodology was utilized to create 500-element SVM input vectors.

The first 5 runs were based on a combination of features extracted from the keyframes of each shot and the whole shot. An overview of the combinations of features used and the shot descriptions that were created is depicted in Figure 1. More information on the extraction and the use of combinations of SIFT descriptors and motion information (interest point tracks) can be found in [12].

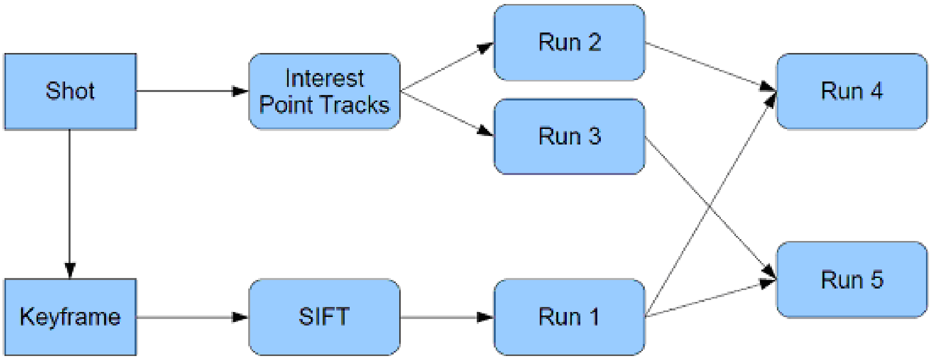


Figure 1: An overview of the combination scheme used to develop the descriptors for each run.

2.3 Results

The runs described above were submitted for the 2009 TRECVID HLFE competition. The evaluation results of the aforementioned runs are given in terms of the Mean Average Precision (MAP) both per run and per high level feature. Table 1 summarizes the results for each run presenting the MAP and the number of true shot predictions.

The SIFT run (ITI-CERTH run 1) was the baseline run of the submission. ITI-CERTH runs 2 and 3 (MSIFT, MTD) performed consistently worse than ITI-CERTH run 1. This can be attributed to the way that the selected sub-set of identified interest point tracks, used in these two runs, was formed. More specifically, the process for generating interest point tracks in a shot, as implemented in these experiments, resulted in the number of tracks identified per shot being at least one order of magnitude higher than the average number of interest points identified in a keyframe. As a result, a small subset of interest point tracks was used for generating the BoW model and descriptors. The track selection process used the estimated reliability of the extracted tracks as a selection criterion, thus not ensuring that the selected subset of tracks adequately described the entire content of the shot (particularly with respect to the distribution of the tracks on the 2D image grid). Furthermore interest point track extraction and use was affected in these preliminary experiments by parameter choices that where later shown to be significantly sub-optimal.

ITI-CERTH runs 4 and 5 (SMS, SMT), on the other hand, performed consistently better that run 1. This can be attributed to the combination of the BoW descriptors of runs 2 and 3, respectively, which sufficiently describe the shot contents along the temporal dimension, with the traditional SIFT-based BoW of a keyframe that sufficiently describes the image information along the spatial dimensions. Furthermore, runs 4 and 5 are shown to be very similar in overall performance, but present differences when considering each high-level feature separately (Figure 2). In particular, for some action-related high-level features (e.g. "Demonstration or Protest"), run 5 that explicitly incorporates motion information is shown to outperform run 4. It should be noted that the aforementioned parameter choices that affected runs 2 and 3 also adversely affected runs 4 and 5.

The results for MESH-ITI-CERTH run 1 (DCT) are directly comparable (in terms of number of visual words, SVM training process, etc.) to the results of ITI-CERTH run 1 (SIFT). These preliminary results (Figure 2) indicate that DCT is strongly outperformed by SIFT.

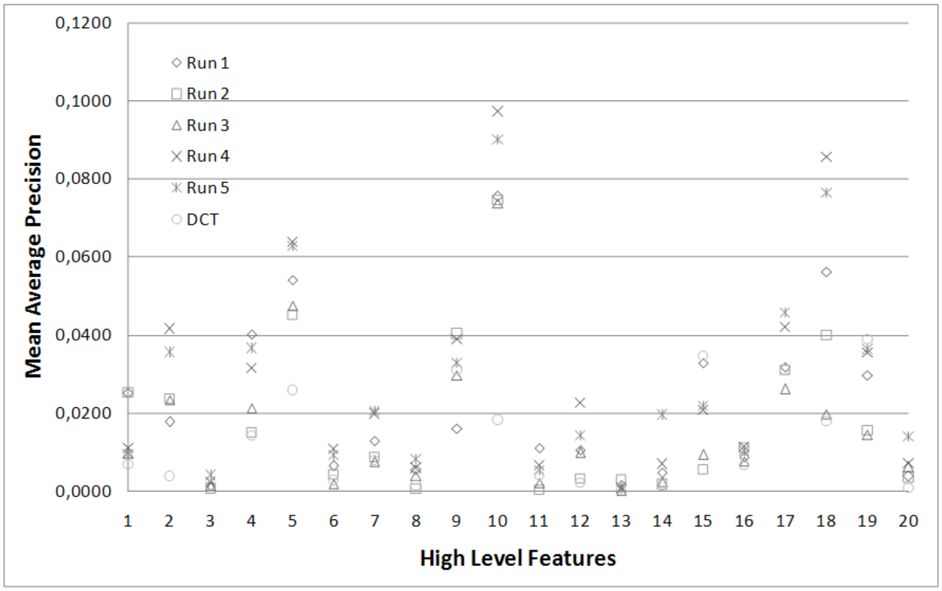


Figure 2: Mean Average Precision per high level feature per run.

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

Run #	ITI-CERTH 1	ITI-CERTH 2	ITI-CERTH 3	ITI-CERTH 4	ITI-CERTH 5	MESH-ITI-CERTH 1
Run name	SIFT	MSIFT	MTD	SMS	SMT	DCT
Total true shots	1749	1438	1426	1874	1885	1347
MAP	0.042	0.032	0.030	0.051	0.050	0.024

3 Interactive Search

3.1 Objectives and Rationale

ITI-CERTH's participation in the TRECVID 2009 search task aimed at studying and drawing conclusions regarding the effectiveness of a set of retrieval modules given the characteristics of the related dataset and the impact of search strategies on system performance. Search strategies were define as guidance provided to the user regarding the functionalities she is allowed to use during a retrieval task in a given order. Within the context of this effort, several runs were submitted, each combining existing modules in a different way, for evaluation purposes while various search strategies were applied in order to study the extent to which they affected overall system performance.

3.2 System Overview

The system used for the search task was an interactive retrieval application that combined basic retrieval functionalities in various modalities (i.e., visual, textual), accessible through a friendly Graphical User Interface (GUI), as shown in Figure 3. The system supported the submission of hybrid queries that combined the available retrieval functionalities, as well as the accumulation of relevant retrieval results, over all queries submitted by a single user during a specified time interval. Using this GUI, the user is allowed to employ any of the supported retrieval functionalities and subsequently filter the derived results using constraints. The retrieval results (representative keyframes of the corresponding shots) were ranked and presented in descending order. The following basic retrieval modules were integrated in the developed search application:

- Visual Similarity Search Module;
- Textual Information Processing and Recommendation Module;
- High Level Concept Retrieval Module;
- High Level Visual and Textual Concepts Fusion Module;

The search system, combining the aforementioned modules, was 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. [13].

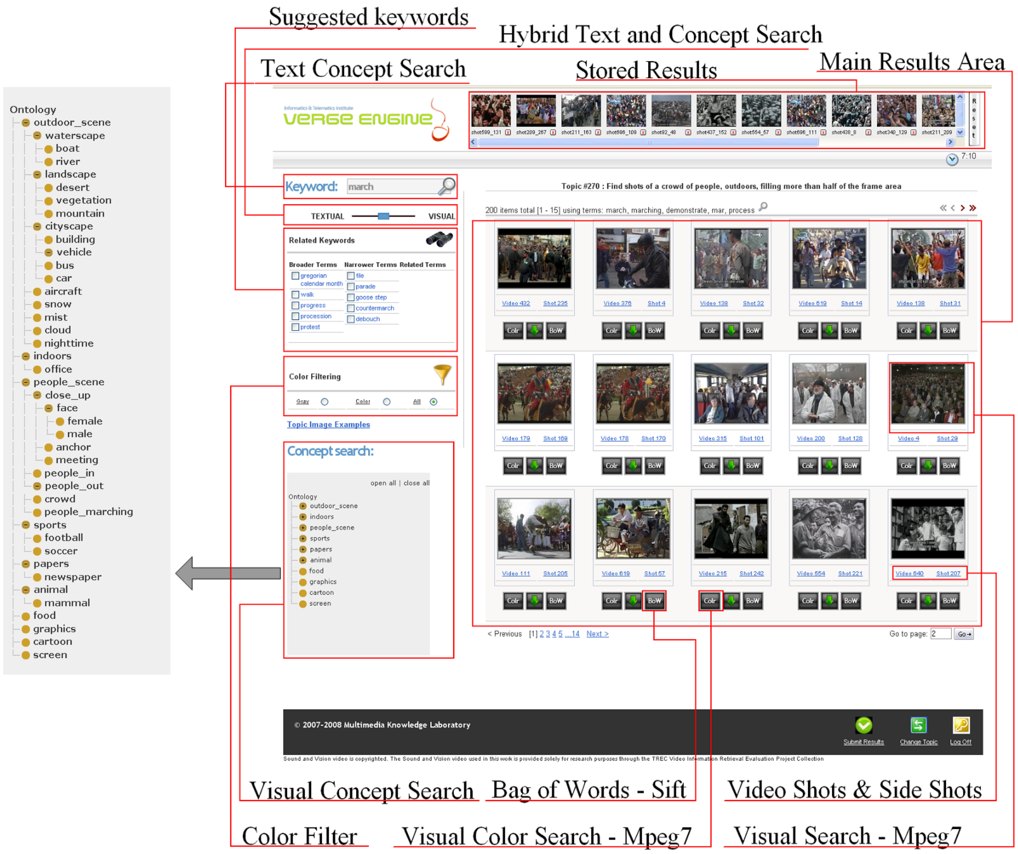
Besides the basic retrieval modules, the system integrated a set of complementary functionalities which aimed at assisting and improving retrieved results. To begin with, the system supported basic temporal queries such as the presentation of neighbouring shots of a specific video shot and the shot-segmented view of each video. Moreover, the system supported a shot preview by rolling three different keyframes. In that way, the user obtains adequate information of the video shot in a very short time. Furthermore, the system had a color filter that constrained results to either grayscale or color images. Finally, the selected shots, considered to be relevant to the query, could be stored by the user. The latter was made possible using a storage structure that mimics the functionality of the shopping cart found in electronic commerce sites and was always visible through the GUI.

A detailed description of each of the aforementioned retrieval modules employed by the system is presented in the following sections.

3.3 Visual similarity search module

The visual search module exploited the visual content of images in the process of retrieving semantically relevant results. Given that the input considered is video, these images were obtained by representing each shot with its temporally middle frame, called the representative keyframe.

In the developed application, the user had two different search options to retrieve visually similar shots. More specifically, content based similarity search was realized using either global information, such as the MPEG-7 visual descriptors that captured different aspects of human perception (i.e., colour and texture), or local information as formulated by vector quantizing the local descriptors obtained by applying the Lowes SIFT transform [9]. Such information was extracted from the representative keyframe of each shot.



Topic #270 : Find shots of a crowd of people, outdoors, filling more than half of the frame area

200 items total [1 - 15] using terms: march, marching, demonstrate, mar, process

© 2007-2008 Multimedia Knowledge Laboratory

Sound and Vision video is copyrighted. The Sound and Vision video used in this work is provided solely for research purposes through the TREC Video Information Retrieval Evaluation Project Collection

Figure 3: User interface of the interactive search platform and focus on the high level visual concepts.

In the first case, five MPEG-7 descriptors, namely, Color Layout, Color Structure, Scalable Color, Edge Histogram, and Homogeneous Texture, were extracted from each image of the collection [14] and stored in a relational database. By concatenating these descriptors, a feature vector was formulated to compactly represent each image in multidimensional space. An empirical evaluation of the system's performance, using different combinations of the aforementioned descriptors, advocated the choice of two MPEG-7 based schemes. The first one relied on color and texture (i.e., ColorLayout and EdgeHistogram were concatenated), while the second scheme relied solely on color (i.e., ColorLayout and ColorStructure).

In the second case, where local information was exploited, the method adopted was an implementation of the bag-of-visual-words approach as described in [15]. Specifically, a large amount of local descriptors (training data) was used for learning the visual words, extracted by applying clustering using a fixed number of clusters. Subsequently, each image was described by a single feature vector that corresponded to a histogram generated by assigning each key-point of the image to the cluster with the closest center. The number of selected clusters (100 in our case) determined the number of bins and, as a consequence, the dimensionality of the resulting feature vectors.

For both cases retrieval was performed efficiently using a multi-dimensional indexing structure. More specifically, an r-tree structure was constructed off-line using the feature vectors of all images and their corresponding image identifiers. R-tree(s) [16] are structures suitable for indexing multi-dimensional objects and known to facilitate fast and efficient retrieval on a large scale. Principal Component Analysis (PCA) [17] was also employed to reduce the dimensionality of the initial space. In the query phase a feature vector was extracted from the query image and submitted to the index structure. The set of resulting numbers corresponded to the identifiers of the images that are found to resemble the query one. Since the order of these identifiers is not ranked according to their level of similarity with the query example, an additional step for ranking these images, using custom distance metrics between their feature vectors, was further applied to yield the final retrieval outcome.

3.4 Textual information processing and recommendation module

The textual query module attempts to exploit the shot audio information using the Machine Translation (MT) output (Dutch to English) for the test video data set provided by [18], which was produced using Automatic Speech Recognition (ASR) output from [19] as a starting point. The resulting annotations were used to create a full-text index utilizing Lemur [13], a toolkit designed to facilitate research in language modeling. The standard pipeline of text processing techniques was used in creating the index. First stopwords were removed from the annotations followed by the application of the Porter [20] algorithm.

Pre-coordination [21] of the Lemur index (by identifying all concepts off-line) was not attempted due to the topical breadth of the corpus (causing a proliferation of term ambiguity), as well as a desire to preserve the "multidimensionality of relationship" among the terms [22]. Instead, post-coordination abilities were enhanced by providing controlled vocabulary terms to the end-user as query term suggestions as well as automatic query expansion terms.

To assist the user in query iteration tasks, a hierarchical navigation menu of suggested keywords was generated from each query submission. Conventional term relationships [23] were supplied by a local WordNet database [24], whereby hypernyms were mapped to broader terms and hyponyms to narrower terms. Related terms were provided by synset terms that were not used for automatic query expansion. This allowed users to browse a concept hierarchy whose taxonomy was determined by the WordNet typology.

All standard post-coordination techniques were implemented for query formulation and editing tasks, including logical Boolean operators (e.g., OR, AND, and NOT), term grouping (using parentheses), and phrases (delimited by quotes), thus allowing the user to manipulate expanded query terms to improve precision.

Recall was boosted by automatic query expansion, also based on WordNet, whereby a list of expanded terms were generated from WordNet synsets. The system was constrained to using the first term in the query string due to the observation that short queries often produce more effective expansions [25, 26]. Since polysemous words tend to degrade precision [27], only the first sense for each query term was considered. Furthermore, subsequent senses often represent marginal usage and since the corpus consisted of general news broadcasts, we expected to predominantly encounter the most common and colloquial sense of each term.

Terms were chosen for expansion by measuring the semantic similarity between each synset term and the original (initial) query term by utilizing the extended gloss overlap method [28, 29]. This technique uses the number of common words found between two term definitions as an indicator of their similarity. Examination of gloss overlaps was confined to synset definitions (rather than the definitions of hypernyms, hyponyms, etc.) due to the observations of [30], who found this relationship to be most effective in significantly and consistently improving retrieval performance. Indeed, it was found that synset overlaps even exceeded other similarity functions, such as mutual information. A threshold was set for expansion termination so that the submitted query was truncated at five terms. Retrieval scores for items containing expanded query terms (e.g., "automobile") were boosted by a factor of 1.5.

Performance of the module in terms of time-efficiency was not satisfactory. The module required between four to ten seconds to return results due to the computation of semantic similarity scores. However, improvements can be gained by moving this functionality off-line. A fairly complete information retrieval thesaurus could be compiled by computing similarity scores for each term in the corpus following the indexing stage. This would eliminate not only the need to invocate the semantic similarity function for many, if not most, queries, but interfacing with the WordNet database could be minimized as well.

3.5 High level visual concept retrieval module

This module facilitated search using high level visual concepts (e.g. water, aircraft, landscape, crowd, etc.). The concepts integrated in the system (Table 2) were selected from the 101 concept list of the MediaMill Challenge Set [31] based on our experience from previous TRECVID workshops as far as performance is concerned. The highlighted concepts of Table 2 are those that we found to exhibit the highest performance. They were depicted in the GUI in the form of a hierarchy (Figure 3).

Table 2: High Level Visual Concepts integrated to the system. The highlighted concepts are depicted in the GUI as suggestions in an ontology tree structure.

aircraft	charts	food	marching_people	people	studio
anchor	cloud	football	meeting	racing	swimmingpool
animal	court	golf	military	river	text
bicycle	crowd	leader_government	mist	road	tower
boat	cycling	graphics	mountain	screen	tree
building	desert	indoor	newspaper	sky	urban
bus	drawing_cartoon	leader_religious	nighttime	smoke	vegetation
car	face	male	office	soccer	vehicle
cartoon	female	mammal	outdoor	snow	walking_running
chair	fish	maps	paper	sports	water

Then we proceeded by extracting high level information (i.e. DoC for the aforementioned concepts) from the test data employing the following approach. A set of MPEG-7-based features were concatenated to form a single "MPEG-7" feature, namely, color structure, color layout, edge histogram, homogeneous texture and scalable color. Also, a SIFT-based feature, that captures the global image features, was created in a 2-stage procedure. A set of 500 keypoints, on average, was extracted from each image. SIFT descriptor vectors (128 elements) were computed for those keypoints, and a set of 100 visual words was created by clustering the multidimensional space that had been created. Using the "Bag of Words" (BoW) methodology [15], a new feature was created from the histogram of distances from each visual word. The MPEG-7 feature was exploited independently from the BoW feature. A set of SVM classifiers (LIBSVM [11] was employed in this case) was used to create classification models for the MPEG-7 and BoW features, using the first half of the development set for the training procedure. The annotation data for the 101 concepts in the training procedure was provided by MediaMill and refer to the TRECVID 2005 dataset. In order to keep the training set balanced between positive and negative input samples, a negative to positive ratio of 5 was kept for all concepts. The SVM parameters were set by an unsupervised optimization procedure, embedded in LIBSVM. The output of the classification is the Degree of Confidence (DoC), by which the query may be classified in a particular concept. After the classification process for both MPEG-7 and BoW features was completed, the respective DoCs were used as input to a second SVM. This stage-2 classification used for training the second half of the development set and self-optimized parameters to create a classification model. The results from this 2-stage SVM on the testing set were sorted by DoC and the similarity co-efficients of each concept were inserted in the database to support the concept retrieval.

Although the procedure described resembles the procedure followed in the HFLE task, it is not identical. The main differences were the type of descriptors used and the number of stages required during the training procedure. The selection of the aforementioned technique for the high level concept module was purely a matter of time since it was already implemented for previous participation whereas the technique developed for TRECVID 2009 HLFE task was implemented concurrently with the search task. However, it is expected that in the following TRECVID conferences the high level module of the search task will integrate the TRECVID 2009 HLFE technique due to its better performance.

3.6 High level visual and textual concepts fusion module

This module combines the high level textual and visual concepts of the aforementioned modules based on a user assisted linear fusion.

From a usability point of view, the user provided a keyword and specified, through a slider, the significance of the textual and visual results by assigning weights. Afterwards, the text recommendation module provided the recommended (synonym, broader, narrower and related) terms, relevant shots and the similarity score of the shots for the specified keyword. Then, the keyword was compared with the list of visual concepts (Table 2) that were integrated in the system and the first matching concept is selected. In parallel, the Degree of Confidence (DoC) of all images for the specific concept were returned. In case no matching concept was found, the recommended terms were compared consecutively with the same list. If a matching concept still could not be found, the visual module provided no input for the specific search. Subsequently, the DoC and the similarity scores of the shots were re-estimated, based on the weights assigned, and the resulting values were added to create the final similarity value.

The procedure is reflected in Equation 1, where i is a specific shot, Sim_i is the final similarity score after the fusion, DoC_i is the degree of confidence, $Score_i$ is the similarity score and, finally, α and β are the weights assigned to their original values, respectively. It should be noted that the number provided by the similarity score and the DoC ranges from 0 to 1, where the higher a value is, the more relevant it is and as it approaches 0, it becomes less relevant.

$$Sim_i = \alpha \cdot DoC_i + \beta \cdot Score_i \text{ where } \alpha+\beta = 1 \quad (1)$$

For instance, if the weight is set to 0.75, the visual module is favored over the textual module since the confidence of the shots returned are multiplied with 0.75 in contrast with those returned by the textual mode, which are multiplied with 0.25. In case the weight is set to zero or one, the results obtained are either exclusively textual or visual.

3.7 Results

We submitted four runs to the Search task. These runs employed different combinations of the existing modules and different searching strategies as described below:

- Run 1: The user was limited to using only visual search (visual similarity search module) since neither the text search nor the visual concepts were supported. As far as the search strategies are concerned, the user was allowed to start the search for each topic using only visual examples.
- Run 2: The user could use visual search and text search but not visual concepts search. As far as the search strategies are concerned, the user could start the search using only text.
- Run 3: The user was allowed to use all the modules integrated into the system (visual, concept, text and concept fusion). Regarding the search strategies, the user could start searching using only visual concepts and text.
- Run 4: The user had at her disposal all the modules (i.e., visual, concept, text and fusion modules) and could apply any search strategies (i.e., visual examples, text and high level visual concepts).

It mush be mentioned that complementary functionalities were available in all runs while the text search also included the results from the term recommendation module. The results achieved using these runs are illustrated in Table 3 below.

The results achieved using these runs are illustrated in Table 3 below.

By comparing the values of Table 3, we can draw conclusions regarding the effectiveness of each of the aforementioned modules and the role of starting strategies in the system's performance. By comparing runs 1, 2 and 3, it is obvious that the high level concepts module improved the results of the system whereas the text search failed to improve the results of run 2 as the performance slightly drops. In that case, it seems that the text search module was incapable of providing good results due to annotations that were produced using automatic machine recording and translation and, consequently, were not sufficiently descriptive of the video shot content. Regarding runs 1, 2 and 3, and considering the different searching strategies applied, we can assume that the use of the visual examples and visual concepts is a better practise. As far as run 4 is concerned, it seems that it has the lowest mean average precision compared to the other runs, although it combines all the possible search methods. The low performance reported in the last run is a consequence of the fact that users were not very familiar with the system, which provided a plethora of different search options, resulting in confusion. In addition, the lack of search strategies exhibited by users in this task seems to have played an important role, as in the previous tasks these strategies could be considered a good guide for the inexperienced user.

In all four runs, the limitation of time was considered to be 10 minutes as defined in this year guidelines.

4 Conclusion

In this paper we reported the ITI-CERTH framework for the TRECVID 2009 evaluation. ITI-CERTH participated in the high-level feature extraction and search tasks. We enhanced the developed system from various aspects. The submissions to all these tasks were successful and employed applications have shown good performance.

Regarding the TRECVID 2009 HLFE competition, many new high level features have been included that have a significant motion pattern. In order to take advantage of the motion activity in each shot we have to employ all frames of the video and not only representative keyframes. Our submission introduced two new features that are based on motion to assist the high level feature representation.

The use of these descriptors, in conjunction with the successful still image representation that SIFT descriptors offer, seems to have an important added value. The evaluation results show a 20% improvement of SMS and SMT over SIFT. It must be noted that the fine tuning of the employed approach for the extraction of the interest point tracks is expected to significantly increase their contribution to improved performance over the baseline SIFT-based approach.

The comparison of DCT with SIFT descriptors showed that there is a significant loss in performance. Nevertheless, the results of using compressed video features could most likely be significantly improved by adopting techniques known for improving SIFT-based BoW description for the purpose of high-level feature extraction, such as examining more than one key-frame per shot, using a pyramidal decomposition for better describing the spatial arrangement of the features, selecting specific MBs based on criteria similar to those used by interest point detectors rather than using all MBs of a frame, etc. Application of any of these techniques on the compressed video information, as suggested in this work, would not affect the faster-than-real-time capabilities of the detection process.

Regarding the TRECVID 2009 search task, a new fusion module was implemented in order to allow the combination of concepts extracted by the textual and visual information. Furthermore, based on runs performed, we drew some general conclusions regarding the leverage of the modules and strategies used on the system's performance. More specifically, the use of high level visual concepts improved the results of the system whereas the text search failed to improve them probably due to the fact that the use of automatic machine recording and translation resulted to low quality descriptions of the video shot content. Finally, regarding the role of searching strategies, the use of visual examples had a positive impact on the system's performance and generally the use of any type of searching strategy assisted the user by providing guidance and avoiding confusion especially in the case of non expert systems.

Table 3: Evaluation of search task results.

Run IDs	run1	run2	run3	run4
Mean Average Precision	0.049	0.046	0.054	0.028

5 Acknowledgements

This work was supported by the projects VIDI-Video (FP6-045547) and CHORUS (FP6-045480), both funded by the European Commission.

References

[1] J. Calic, P. Kramer, U. Naci, and et. al S. Vrochidis. Cost292 experimental framework for trecvid 2006. 4th TRECVID Workshop, Gaithersburg, USA, November 2006, 2007.

[2] Q. Zhang, K. Chandramouli, U. Damnjanovic, T. Piatrik, and E. Izquierdo et al. The cost292 experimental framework for trecvid 2007. 5th TRECVID Workshop, Gaithersburg, USA, November 2007, 2007.

[3] Q. Zhang, G. Tolias, B. Mansencal, and A. Saracoglu et al. Cost292 experimental framework for trecvid 2008. 6th TRECVID Workshop, Gaithersburg, USA, November 2008, 2008.

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

[5] 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.

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

[7] P. Wilkins, T. Adamek, and G. J.F.Jones et al. K-space at trecvid 2007. 5th TRECVID Workshop, Gaithersburg, USA, November 2007, 2007.

[8] P. Wilkins, D. Byrne, H. Lee, and K. McGuinness et al. K-space at trecvid 2008. 6th TRECVID Workshop, Gaithersburg, USA, November 2008, 2008.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[10] K. E. A. van de Sande and T. Gevers and C. G. M. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (in press), 2010.

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

[12] V. Mezaris, A. Dimou, and I. Kompatsiaris. Local Invariant Feature Tracks for High-Level Video Feature Extraction. In *Proceedings of the 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2010)*, Desenzano del Garda, Italy, April 2010.

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

[14] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, Mar 1998.

[15] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. volume 2, pages 1470–1477, 2003.

[16] A. Guttman. R-trees: a dynamic index structure for spatial searching. In *SIGMOD '84: Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, pages 47–57, New York, NY, USA, 1984. ACM.

[17] I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, NY, USA, 2002.

[18] S. Carter, C. Monz, and S. Yahyaei. The QMUL System Description for IWSLT 2008. In *Proc. of the International Workshop on Spoken Language Translation*, pages 104–107, Hawaii, USA, 2008.

[19] M.A.H. Huijbregts, R.J.F. Ordelman, and F.M.G. de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of the Second International Conference on Semantic and Digital Media Technologies, SAMT 2007*, volume 4816 of *Lecture Notes in Computer Science*, pages 78–90, Berlin, December 2007. Springer Verlag.

[20] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[21] J. Aitchison, A. Gilchrist, and D. Bawden. Europa Publications Ltd, 4th edition edition, 2000.

[22] F. W. Lancaster. *Vocabulary control for information retrieval.* Information Resources Press, Arlington, Virginia, 2nd edition edition, 1986.

[23] Guidelines for the construction, format, and management of monolingual controlled vocabularies, 2005. `http://www.niso.org/standards`.

[24] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication).* The MIT Press, May 1998.

[25] E. M. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[26] H. Fang and C. X. Zhai. Semantic term matching in axiomatic approaches to information retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122, New York, NY, USA, 2006. ACM.

[27] M. Rila, T. Takenobu, and T. Hozumi. The use of wordnet in information retrieval, 1998.

[28] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, 2003.

[29] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, NY, USA, 1986. ACM.

[30] H. Fang. A Re-examination of Query Expansion Using Lexical Resources. In *Proceedings of ACL-08: HLT*, pages 139–147, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[31] C. G. M. Snoek, M. Worring, J. C. van Gemert, J. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, New York, NY, USA, 2006. ACM.