ITI-CERTH participation to TRECVID 2010

Anastasia Moumtzidou, Anastasios Dimou, Nikolaos Gkalelis,
Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris

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, gkalelis, stefanos, bmezaris, ikom}@iti.gr

October 22, 2010

Abstract

This paper provides an overview of the tasks submitted to TRECVID 2010 by ITI-CERTH. ITI-CERTH participated in the Known-item search (KIS) and Instance search (INS) tasks, 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 SIFT and Bag-of-Words for shot representation. In the MED task, trained concept detectors are used to represent video sources with model vector sequences, while a dimensionality reduction method is used to derive a discriminant subspace for recognizing events. The KIS and INS search tasks are performed with by employing VERGE, which is an interactive retrieval application combining retrieval functionalities in various modalities (i.e. textual, visual and concept search). Evaluation results on the submitted runs for the aforementioned tasks provide interesting conclusions regarding the performance of the involved techniques and algorithms.

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 [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 [2, 3, 4]) and in the semantic indexing (SIN) task (which is the similar to the old high-level feature extraction task) under MESH integrated project [5] (TRECVID 2008 [6]), K-SPACE project [7] (TRECVID 2007 and 2008 [8, 9]). Recently, ITI-CERTH has participated as stand alone organization in the search and high level feature tasks of TRECVID 2009 [10]. 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 four tasks: known-item search, instance search, semantic indexing and the event detection in internet multimedia tasks. Regarding the event detection (MED) task, it should be noted that is the first year that ITI-CERTH participates in this task. In the following sections we will present in detail the applied algorithms and the evaluation for the runs we performed in the aforementioned tasks.

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 keyframes only (e.g. [6]). The motion information of

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

the shot, particularly local (object) motion, is vital when considering action-related concepts. Such concepts are also present in TRECVID 2010 SIN task (e.g. ”Swimming”, ”Walking”, ”Car racing”). In TRECVID 2010, ITI-CERTH examines how feature tracks, which are extracted from the processing of the entire shot instead of selected keyframes and are employed in a Bag-of-Spatiotemporal-Words (BoSW) strategy [14], can be combined with performance enhancing techniques, such as spatial/temporal pyramidal decomposition, to lead to improved high-level feature extraction in video. Moreover, a re-ranking scheme has been tested, based on the quality of the video. 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 feature tracks can extend the traditional SIFT-based Bag-of-Words (BoW) technique. All 4 runs were based on generating one or more Bag-of-Words models, using feature tracks that jointly capture motion and 2D appearance information in each shot and/or SIFT descriptors that capture only 2D appearance (i.e. intensity distribution in a local neighborhood). SIFT descriptors were extracted from an image using the original 128-dimensional SIFT descriptors introduced in [11], utilizing the software implementation of [12]. In all cases where a BoW or BoSW model was defined, the number of words was set to 500. In order to enhance the performance, spatial and temporal pyramidal decomposition was used. For the BoW descriptor, a two-level spatial pyramidal scheme was employed for every keyframe. The first level is decomposed in a 2x2 scheme and the second in a 3x1, to create a concatenated keyframe description vector of dimension 4000. For the feature tracks-based BoSW model a one-level temporal pyramidal scheme was used. Each shot was split into 3 equal time slots and a BoSW model was created for each one of them. As a result, a concatenated description vector of dimension 2000 was extracted for each shot.

A common training 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. Unlike the 2009 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 30000 training samples per concept. A variable proportion of positive/negative samples was used to reach the 30000 samples limit for as many concepts as possible; this proportion ranged from 5:1 to 1:1. The SVM parameters were set using an unsupervised optimization procedure that is part of the LIBSVM tool [13]. 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 results per high-level feature were sorted by DoC in descending order and the first 2000 shots were submitted to NIST.

In two of the submitted runs, a re-ranking scheme was also tested. The adopted re-ranking methodology relies on a very simple estimation of the video quality of each shot, specifically on examining the number of interest points detected by the SIFT descriptor extractor the keyframes of the shot.

The 4 submitted runs were:

- ITI-CERTH-Run 1: “BoSW + BoW”. This is a run combining the traditional BoW model, which is created from the SIFT features extracted from one representative keyframe of the shot, with the BoSW model based on feature tracks extracted from the entire shot. Spatial pyramidal decomposition was used for the SIFT features and temporal pyramidal decomposition was used for the feature tracks. The combined descriptor has a total dimension of 6000.

- ITI-CERTH-Run 2: “BoSW + BoW Re-Ranked”. In this run, the results of run 1 are re-ranked using the aforementioned video quality-based approach.

- ITI-CERTH-Run 3: “BoW”. This is a baseline run using the traditional Bag-of-Words model, which is created from the SIFT features extracted from one representative keyframe of the shot Soft-Binning and spatial pyramidal decomposition was used.

- ITI-CERTH-Run 4: “BoW Re-Ranked”. In this run, the results of run 3 are re-ranked using the aforementioned video quality-based approach.

More information on the extraction and the use of combinations of SIFT descriptors and motion information (feature tracks) can be found in [14].

2.3 Results

The runs described above were submitted for the 2010 TRECVID SIN competition. The ITI-CERTH participation topped in performance among the runs submitted from all participating institutions in concepts ”Walking and ”Sitting down”. 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 MAP and the number of true shot predictions.

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**
Run name	BoW+BoSW	BoW+BoSW Re-ranked	BoW	BoW Re-ranked
Total true shots	7972	7972	7340	7340
MxinfAP	**0.038**	**0.035**	**0.030**	**0.028**

The BoW run (ITI-CERTH run 3) was the baseline run of the submission. It combines the traditional SIFT-based bag-of-words model with soft-binning and spatial pyramidal decomposition to establish a baseline but robust performance run. The BoW Re-ranked run (ITI-CERTH run 4) is used to assess a new re-ranking technique based on a very simple estimation of objective video quality. This technique did not show any performance gain (Figure 1), performing overall worse than the baseline run.

ITI-CERTH run 1 incorporates motion information (feature tracks) that is extracted from the whole shot in addition to the baseline run. It performed significantly better than the baseline ITI-CERTH run 3, performing better for most of the concepts (Figure 1). This is due to the combination of the BoW descriptors with the wealth of motion information that the feature tracks contain. Performance gain can be observed for both static and action-related high-level features. For the static high-level features (e.g. ”Cityscape”) the gain can be attributed to the mean-SIFT part of the feature tracks descriptor that produces a more robust description of the local intensity distribution, due to averaging. Performance gain is significantly higher for action-related high-level features (e.g. ”Swimming”, ”Walking”).

The BoW+BoSW Reranked run (ITI-CERTH run 2) is using the results from ITI-CERTH run 1 with the reranking technique described above. Re-ranking performed again worse than the original run 1.

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 requiring complex and computationally expensive algorithms. For applications that require near real-time video processing, e.g., video indexing and retrieval systems, the training and especially the testing time of the recognition algorithms is a very critical quality factor. The objective of our participation in MED is to evaluate the recognition performance for a number of efficient algorithms that we have recently developed, and compare them with current state of the art algorithms.

In these experiments we use the “model vector-based approach” presented in [15] for recognizing the three events defined in the 2010 TRECVID MED evaluation, namely, “making a cake”, “batting a run”, and “assembling a shelter”. This method is described briefly in the following: Automatic techniques are initially used for the temporal segmentation of videos to shots [16] and their description with low-level visual features. Then, each shot is represented with a model vector which comprises the responses of 231 trained concept detectors. The concepts used to this end are the 101 MediaMill

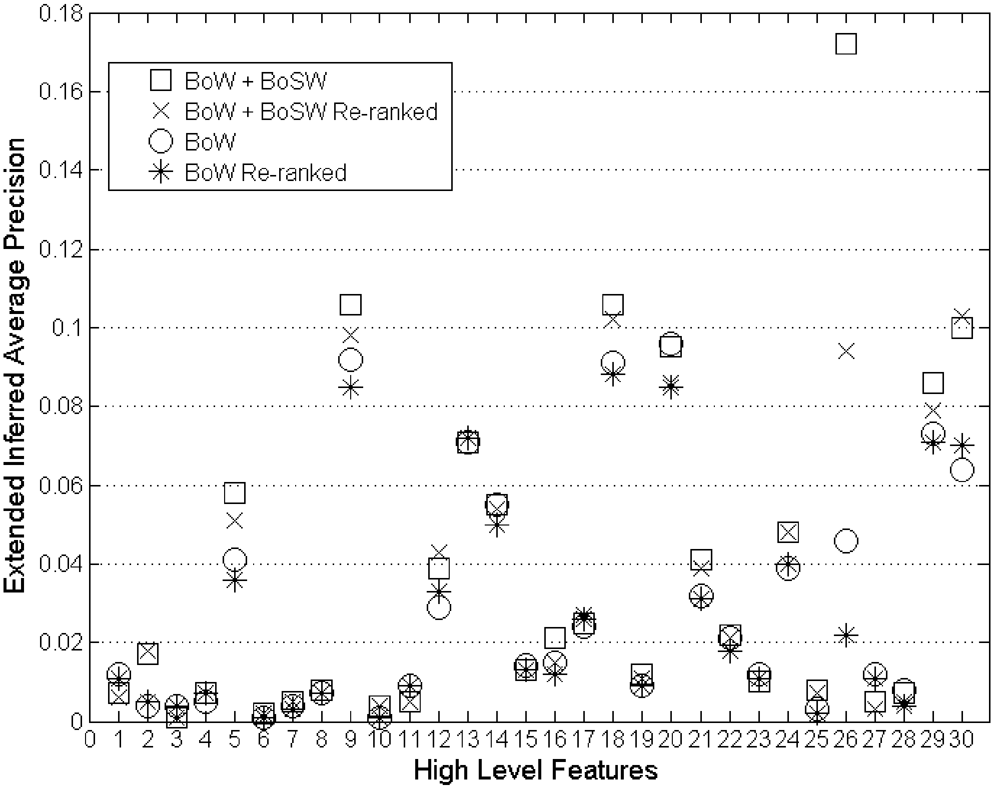


Figure 1: Extended Inferred Average Precision per high level feature per run.

Challenge concepts and the 130 TRECVID 2010 SIN Task concepts. The respective annotated datasets have been used to train the 231 concept detectors as described in [15, 10]. The confidence values derived from each trained concept detector are concatenated to form a 231-dimensional model vector. Thus, a model vector is used to represent a video shot. These model vectors are used as input to a subspace technique for obtaining the respective projection matrix [17, 18]. During testing the projection matrix is used to project the model vectors of the test shots to the discriminant subspace and the nearest neighbor classifier (NN) along with the Hausdorff distance are used to recognize an event.

The use of concept detectors trained on pre-existing datasets (MediaMill, TRECVID SIN Task) for forming model vectors can significantly reduce the training time as no additional time is introduced for training event and/or concept detectors specifically for the dataset of interest. It is expected that the event recognition performance of such an approach will depend on the relevance of the concept detectors with the investigated event as well as on the total number of available trained concept detectors. Additionally, the utilization of a subspace technique is expected to provide a performance gain in both the recognition accuracy and testing time. Specifically, one of the most popular dimensionality reduction methods is linear discriminant analysis (LDA). However, LDA has many restrictions, mainly requiring that the classes are linearly separable, which usually is not true in real-world applications. For solving this problem, Kernel LDA (KLDA) was designed for separating non-linearly distributed classes. However, the flexibility of KLDA comes at the cost of a very high testing time that scales with the number of training samples. Recently, SDA has been proposed, which produces a projection matrix to separate piece-wise separable data and, thus, in comparison to KLDA, provides much faster testing times. More recently, we developed a number of SDA variants, generally called improved SDA (ISDA) in the sequel to further improve the recognition accuracy and the degree of dimensionality reduction achieved by SDA.

Another common problem in various event detection tasks is the classification of asymmetric classes. In such tasks, the "positive" classes contain the single events that need to be detected and the "negative" class is "the rest of the world" class containing all other samples [19]. Inspired from this fact, we have developed an asymmetric version of the SDA for alleviating this problem.

Upon this analysis the objectives of the submission are:

- To compare our ISDA method and its variants with the SDA method and simpler approaches (LDA, or processing the model vectors directly in the input space) in both recognition accuracy and degree of dimensionality reduction.
- To compare our general approach, which relies on model vectors and does not use an elaborate training method for learning the events from an event-annotated dataset, with competitive runs submitted by other institutions participating to the MED task.

3.2 Description of runs

A common training method for all runs was selected to provide comparable results between them[6, 15]. In particular, a Bag-of-Words (BoW) procedure is used to represent each shot keyframe with a high-dimensional feature vector. This is done by firstly extracting keypoints from each keyframe and using them to form 128-dimensional SIFT [11] vectors, secondly clustering the SIFT vectors to create a number of Visual Words in the 128-dimensional feature space, and thirdly using the created Visual Words for mapping the 128-dimensional feature vectors to the new high-dimensional feature space. This feature vector is then used as input to each one of the 231 SVM-based concepts detectors. 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. The model vectors described above are used for optimizing the parameters of the employed subspace algorithm (e.g., dimensionality of output vectors) as discussed below. The optimization process was guided by the Normalized Detection Cost (NDC), i.e., NDC was the quantity to be minimized. During testing, the same procedure is followed to represent each test shot with the respective model vector. The model vector is further projected in the discriminant subspace using the corresponding optimized subspace algorithm, and is classified in this space using the NN rule together with a variant of the Hausdorff distance, depending on the run. The nine submitted runs are described in the following:

- ITI-CERTH_2010_MED_EVAL_c-IN_1: This is the baseline run, i.e, the Hausdorff distance is used to compare the model vector sequences directly in the input space of dimensionality $D = 231$.
- ITI-CERTH_2010_MED_EVAL_c-LDA_1: In this run we use conventional LDA to project the model vectors to a discriminant subspace. The dimensionality of the vectors in the discriminant subspace is $D = 6$.
- ITI-CERTH_2010_MED_EVAL_c-SDA_1: This run uses the original SDA. The dimensionality of the vectors in the discriminant subspace is $D = 90$.
- ITI-CERTH_2010_MED_EVAL_p-ISDA1_1: In this run we use the improved SDA (ISDA) method developed at ITI-CERTH. The dimensionality of the model vectors in the discriminant subspace is $D = 42$.
- ITI-CERTH_2010_MED_EVAL_c-ISDA2_1: In this run we use the ISDA algorithm to produce a confidence score for each event and for each video in the development set. These confidence scores are used to build a Gaussian distribution confidence score model for each event. During testing the ISDA algorithm produces three confidence scores, one for each event. Then these scores are weighted with the probability that these scores belong to the respective event using the event probability models.
- ITI-CERTH_2010_MED_EVAL_c-ISDA3_1: This is also the ISDA algorithm but in this case the detection of an event in the discriminant subspace is done using a windowing version of the Hausdorff distance.

Table 2: Evaluation results for event “batting in run”.

Run	c-IN	c-LDA	c-SDA	p-ISDA1	c-ISDA2	c-ISDA3	c-ISDA4	c-ISDA5	c-ISDA6
TP	32	21	27	29	29	28	29	28	28
FA	48	30	36	35	37	31	32	35	35
NDC	0.6766	0.7766	0.6936	0.6436	0.6584	0.6351	**0.6213**	0.6649	0.6649
D	231	6	90	42	42	42	42	42	42

Table 3: Evaluation results for event “assembling shelter”.

Run	c-IN	c-LDA	c-SDA	p-ISDA1	c-ISDA2	c-ISDA3	c-ISDA4	c-ISDA5	c-ISDA6
TP	13	6	11	9	10	11	11	10	10
FA	52	24	28	38	41	45	49	39	39
NDC	1.1044	1.0482	**0.9692**	1.0871	1.0877	1.0958	1.1255	1.0728	1.0728
D	231	6	90	42	42	42	42	42	42

Table 4: Evaluation results for event “making cake”.

Run	c-IN	c-LDA	c-SDA	p-ISDA1	c-ISDA2	c-ISDA3	c-ISDA4	c-ISDA5	c-ISDA6
TP	12	6	5	6	11	5	8	5	5
FA	36	43	15	15	33	21	59	14	14
NDC	1.0127	1.1925	0.9979	**0.9840**	1.0117	1.05	1.2691	0.9979	0.9979
D	231	6	90	42	42	42	42	42	42

- ITI-CERTH_2010_MED_EVAL_c-ISDA4_1: This run combines the procedures followed in runs c-ISDA2_1 and c-ISDA3_1. That is, model vectors are projected using ISDA, the windowing version of the Hausdorff distance is used to produce a confidence value for each event, and confidence values are weighted with weights resulting from the probability event models build during training.

- ITI-CERTH_2010_MED_EVAL_c-ISDA5_1: This run uses the asymmetric version of the ISDA algorithm. Similarly to ISDA, the dimensionality of the model vectors in the discriminant subspace is $D = 42$.

- ITI-CERTH_2010_MED_EVAL_c-ISDA6_1: In this run the asymmetric version of the ISDA algorithm is used, as in c-ISDA5_1, and additionally, confidence values are weighted with probabilistic weights, using a procedure similar to the one described for c-ISDA2_1.

3.3 Results

The runs described above were submitted for the 2010 TRECVID MED competition. The evaluation results for the events “batting in run”, “assembling shelter” and “making cake” are provided in Tables 2, 3 and 4 respectively. The results are given in terms of the True Positives (TP), False Alarms (FA), Normalized Detection Cost (NDC), and dimensionality of the model vectors in the discriminant subspace (D). From the above analysis we observe the following:

- Although we use a “model vector-based approach” together with a simple NN classifier, i.e., we do not explicitly train event-specific classifiers using directly the provided training set, still very good recognition results are achieved for the event “batting in run” ($NDC = 0.6213$) and moderately good results are achieved for the other two events. The good performance regarding the event “batting in run” can be explained by the fact that several of our trained detectors have been trained to recognize relevant event concepts such as “grass”, “running”, etc. On the other hand, for the other two events, although our set of trained concept detectors does not include directly relevant concepts, a moderately good performance is still achieved. Therefore, we can conclude that our “model vector-based approach” can indeed provide good results even without using an elaborate and computationally-costly learning algorithm.

- The best recognition rates among our runs were achieved by the following algorithms: a) c-ISDA4 for the event “batting in run” ($NDC = 0.6213$) , b) c-SDA for the event “assembling shelter” ($NDC = 0.9692$), and c) p-ISDA1 for the event “making cake” ($NDC = 0.9840$).

- The SDA (c-SDA) and all ISDA methods (p-ISDA1, c-ISDA2, c-ISDA3, c-ISDA4, c-ISDA5, c-ISDA6) performed consistently better for all events, compared to the conventional LDA (c-LDA) as well as to the method that directly classifies the test model vector sequences in the input space (c-IN).

- p-ISDA1 (as well as all ISDA methods) provided better detection rates from SDA regarding the two of the three events (i.e., for the “batting in run” and “making cake” events). In addition, ISDA provided a better degree of dimensionality reduction ($D_{ISDA} = 42$ while $D_{SDA} = 90$), leading to a considerable system speedup especially for large video sequences with many shots. Concluding, we can say that ISDA, compared to SDA, indeed contributes to better performance both in recognition accuracy (NDC) and system speedup.

4 Interactive Search

4.1 Objective of the submission

ITI-CERTH’s participation in the TRECVID 2010 known-item (KIS) and instance (INS) search tasks 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.

In the following, we describe briefly the characteristics of each task.

- Known-item search task (KIS): The KIS task represents the situation, in which the user is searching for a specific video contained in a collection. It is assumed that the user already knows the content of the video, therefore a detailed textual description is provided.

- Instance search task The instance search task focuses on retrieving segments of a certain specific person, object, or place, given a visual example. For each visual example a binary mask is provided that specifies the region of interest/ segment as well as the inner region against a grey background and a list of vertices for the inner region.

4.2 System Overview

The system employed for both search tasks was VERGE, which is an interactive retrieval application that combines basic retrieval functionalities in various modalities (i.e., visual, textual), accessible through a friendly Graphical User Interface (GUI), as shown in Figure 2. The system supports the submission of hybrid queries that combine the available retrieval functionalities, as well as the accumulation of relevant retrieval results by a single user. The following basic retrieval modules are integrated in the developed search application:

- Visual Similarity Search Module;

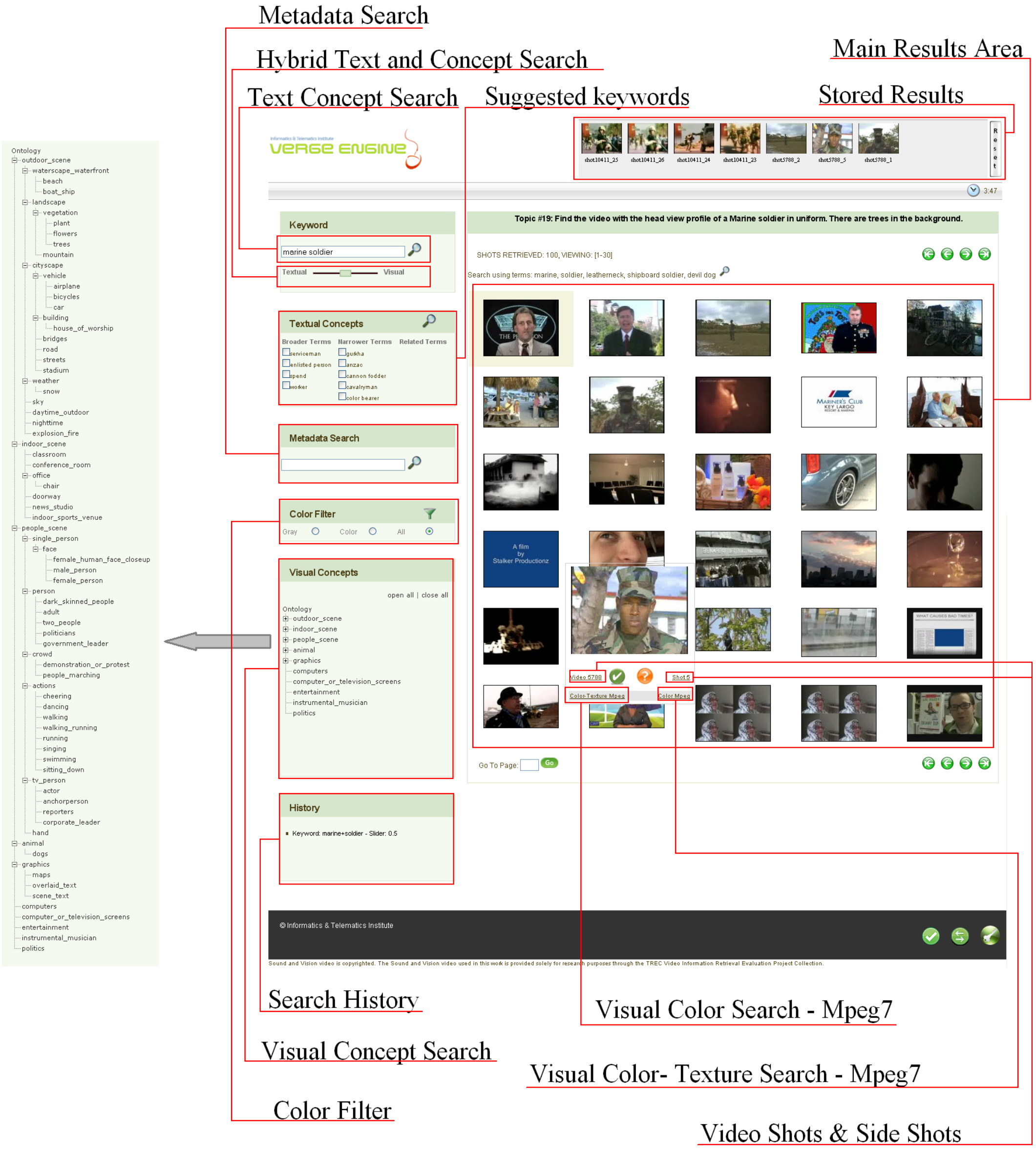
- Textual Information Processing and Recommendation Module;

- Metadata Processing and Retrieval Module;

- High Level Concept Retrieval Module;

- High Level Visual and Textual Concepts Fusion Module;

- Segmentation Module;



Topic #19: Find the video with the head view profile of a Marine soldier in uniform. There are trees in the background.

SHOTS RETRIEVED: 100, VIEWING: [1-30]

Search using terms: marine, soldier, leatherneck, shipboard soldier, devil dog

- Keyword: marine+soldier - Slider: 0.5

© Informatics & Telematics Institute

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

The search system, combining the aforementioned modules, 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 [20].

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 presentation of temporally adjacent shots of a specific video shot and the shot-segmented view of each video. Moreover, the system supports a shot preview by rolling three different keyframes. Furthermore, the system provides a color filter option that constrains the results to either grayscale or color images. 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 new functionality that has been added this year is the history bin, where all the user actions are recorded, which allows for easy navigation to previous results.

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

4.2.1 Visual similarity search module

The visual similarity search module performs image content-based retrieval with a view to retrieving visually similar results. Given that the input considered is video, these images are obtained by representing each shot with its temporally middle frame, called the representative keyframe.

Following the visual similarity module implementation in [10], five MPEG-7 descriptors, namely, Color Layout, Color Structure, Scalable Color, Edge Histogram, and Homogeneous Texture, are extracted from each image of the collection [21]. By concatenating these descriptors, a feature vector is 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 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). Finally, efficient retrieval is achieved by employing multi-dimensional r-tree [22] indexing structure. Principal Component Analysis (PCA) [23] is also employed to reduce the dimensionality of the initial space.

4.2.2 Textual information processing and recommendation module

The textual query module exploits the shot audio information. The implementation of this module is based on the text processing and retrieval techniques presented in [10]. To begin with, Automatic Speech Recognition (ASR) is applied on test video data. Then, in case that the spoken language is not English, a Machine Translation (MT) step is also required. In this implementation, the ASR and the MT are provided by [24, 25, 26] and [27] respectively.

The textual information generated is used to create a full-text index utilizing Lemur [20], a toolkit designed to facilitate research in language modeling.

To assist the user in query iteration tasks, a hierarchical navigation menu of suggested keywords is generated from each query submission. Conventional term relationships [28] are supplied by a local WordNet database [29], whereby hypernyms are mapped to broader terms and hyponyms to narrower terms. Related terms are provided by synset terms that were not used for automatic query expansion. Recall is boosted by automatic query expansion, also based on WordNet, whereby a list of expanded terms were generated from WordNet synsets.

4.2.3 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 IACC.1 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 is indexed into the database for fast retrieval.

We should note that a more complicated processing and recommendation system of the metadata similar to the one developed for the textual information, is expected to be applied in the next TRECVID workshops.

4.2.4 High level visual concept retrieval module

This module facilitates search by indexing the video shots based on high level visual concept information such as water, aircraft, landscape, crowd. The concepts that are extracted, as well as the procedure used for high level visual concept extraction are different for the two search tasks.

As far as the INS task is concerned, the 101 concept list of the MediaMill Challenge Set [30] was used. From the aforementioned list, only 37 of them (e.g. waterscape, boat, cityscape, office) that exhibit the highest performance are selected. These concepts are depicted in the GUI in the form of a hierarchy (Figure 2). Regarding the approach followed for extracting high level visual concepts, a set of MPEG-7-based features is concatenated to form a single “MPEG-7” feature, namely, color structure, color layout, edge histogram, homogeneous texture and scalable color. Then, a set of SVM classifiers (LIBSVM [13]) is used to create classification models for the MPEG-7 features, using the first half of the development set for the training procedure [10].

As far as the KIS task is concerned, we selected the best performing 72 of the 130 concepts that have been selected for the TRECVID 2010 semantic indexing task. Regarding the approach for extracting high level concepts, SIFT descriptors are extracted from a single keyframe for each shot and feature tracks are extracted using all frames of the shot. Using a Bag-of-Words-type of methodology, BoW and BoSW descriptors are produced and their combination is used to train SVM classification models [14].

The annotation data for the concepts used in the training procedure for the INS task is provided by MediaMill and refer to the TRECVID 2005 dataset, while these for the KIS task refer to the SIN task training dataset.

In both tasks, the output of the classification is the Degree of Confidence (DoC), by which the query may be classified in a particular concept. However, the procedure for obtaining these degrees is different for each task. More details for the INS task can be found in the High level visual concept retrieval section of [10] while for KIS task in section 2.

4.2.5 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. 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.

$$Sim_i = \alpha \cdot DoC_i + \beta \cdot Score_i \text{ where } \alpha{+}\beta = 1 \tag{1}$$

4.2.6 Segmentation module

The segmentation module segments an image into segments in order to make possible the comparison between parts of images. In order to extract this information a two-step off-line procedure is realized. During the first step, the keyframes are broken down into semantic objects, while during the second the visual description of the acquired objects is realized. In the following, we describe thoroughly only the first step of the procedure, since the second is similar to the procedure followed in the visual similarity search module.

The segmentation step, which is called large-format image segmentation scheme, is analyzed into two parts [31, 32]. The first part refers to the implementation of a color image segmentation algorithm and it is based on a variant of the K-Means-with-connectivity constraint algorithm (KMCC). This algorithm classifies the pixels into regions taking into account not only the intensity information associated with each pixel but also the position of the pixel, thus producing connected regions that correspond to the real-life objects rather than sets of chromatically similar pixels. This is accomplished through the use of texture features in combination with the intensity and position features; this, along with the texture-dependent filtering of pixel intensities (conditional filtering), endow the segmentation algorithm with the capability to handle textured objects effectively, by forming large, chromatically non-uniform regions instead of breaking down the objects to a large number of chromatically uniform regions, as was the case in a previous preliminary version of the algorithm. In addition, in the proposed algorithm the required initial values are estimated using an initial clustering procedure, based on breaking down the image to square blocks and assigning an intensity feature vector and a texture feature vector to each block. During this phase a variant of the maximin algorithm is applied. The result of the application of the segmentation algorithm to a color image is the segmentation mask, i.e. a grayscale image, in which different gray values correspond to different regions. While the aforementioned segmentation approach is considerably fast when the algorithm is applied to images of relatively small dimensions, its time efficiency degrades quickly as the image size increases. Thus, the second part of the module focuses on shortening the segmentation time. Hence, in order to provide a more efficient scheme for the segmentation of large images regions of image falling below a size threshold could be omitted. However, although this assumption improves the time efficiency of the segmentation process, it degrades the quality of the segmentation result, since edges between objects are crudely approximated by piecewise linear segments, lowering the perceptual quality of the result. To alleviate this problem, the use of the Bayes classifier for the reclassification of pixels belonging to blocks on edges between regions is proposed. At this point, we should note that the two parts are not directly interrelated and thus they can be applied separately.

During the second step of the segmentation module, the visual description of the acquired objects is realized. Both the original image and the segmentation mask are used as input of the MPEG-7 extraction tool (which is also used in the visual similarity search module) and the outcome consists of MPEG-7 descriptors for each region. An overview of the procedure followed is depicted in Figure 3.

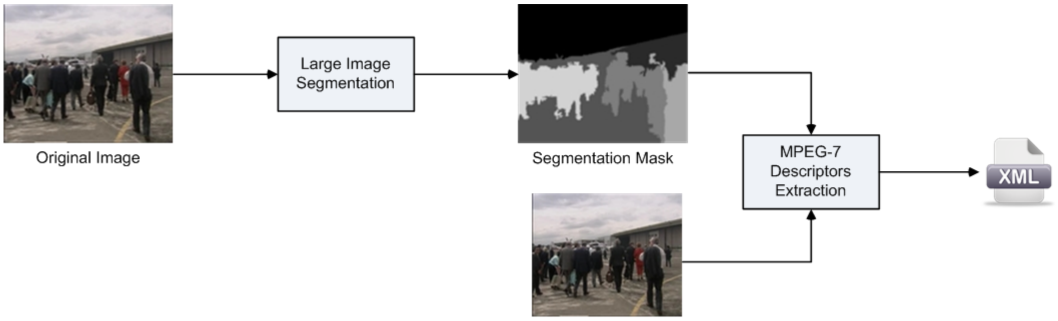


Figure 3: Segmentation module.

4.3 Instance Search Task Results

The system developed for the instance search task includes all the aforementioned modules apart from the metadata as they weren't any available for this task. We have submitted the following two runs:

- I_X_NO_ITI-CERTH_1: The user was allowed to use all the modules integrated into the system (i.e. visual, concept, text, concept fusion and segmentation).
- I_X_NO_ITI-CERTH_2: The user could use the visual, concept, text, concept fusion modules during the search. Therefore, the only module that wasn't on the user's disposal was the segmentation module.

It mush be mentioned that complementary functionalities were available in all runs. In both runs, the limitation of time was considered to be 15 minutes. The results achieved using these runs are illustrated in Table 5 and in Figures 4 and 5.

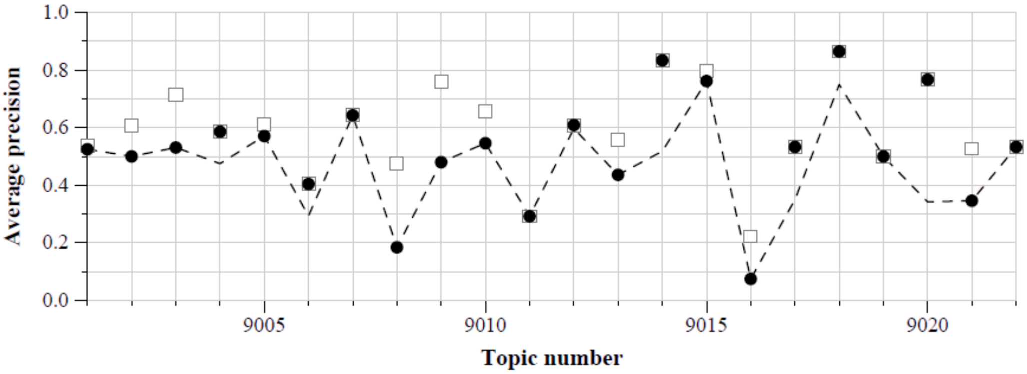


Figure 4: Graph with Average precision of first run of INS task.

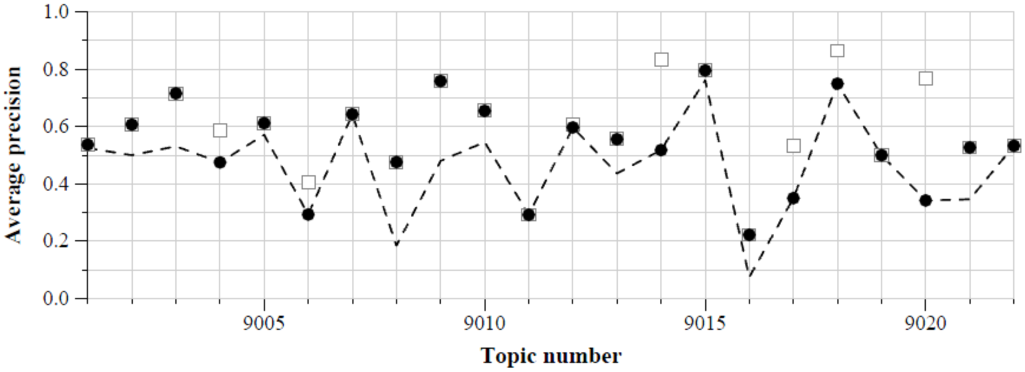


Figure 5: Graph with Average precision of second run of INS task.

By comparing the values of Table 5, we can draw conclusions regarding the effectiveness of the segmentation module. Thus, by comparing the runs we can deduce that the second run has slightly better performance than the first, which includes all the modules. However one could expect a major improvement due to the involvement of the segmentation module. A possible explanation is that the segmentation module was considerable slower that the simple visual search and thus less queries could be executed in the given time. In addition, in many cases the segmentation module was unable to segment the keyframes into semantically meaningful objects and therefore the system was performing comparisons between segments being part of the semantic object and yielding not useful results. Hence, we can conclude that the employment of the specific segmentation module didn't assist as much as expected in this task.

4.4 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 5: Evaluation of search task results.

Run IDs	Mean Average Precision
I_X_NO_ITI-CERTH_1	0.524
I_X_NO_ITI-CERTH_2	0.534

- I_A_YES_ITI-CERTH_1: The user had at her disposal all the modules (i.e., visual, concept, text, fusion and metadata modules).
- I_A_YES_ITI-CERTH_2: The user could use all the modules integrated to the system apart from the high level visual concept and fusion module.
- I_A_NO_ITI-CERTH_3: The user could use all the modules integrated to the system apart from the metadata module.
- I_A_NO_ITI-CERTH_4: The user was allowed to used only the visual and text module.

It should be clarified that no metadata searching is allowed in Runs 3 and 4. The complementary functionalities were available in all runs, while the time duration for each run was considered to be 5 minutes. The results achieved using these runs are illustrated in Table 6 and in Figures 6, 7, 8 and 9.

Table 6: Evaluation of search task results.

Run IDs	Mean Average Precision
I_A_YES_ITI-CERTH_1	0.273
I_A_YES_ITI-CERTH_2	0.409
I_A_YES_ITI-CERTH_3	0.136
I_A_YES_ITI-CERTH_4	0.182

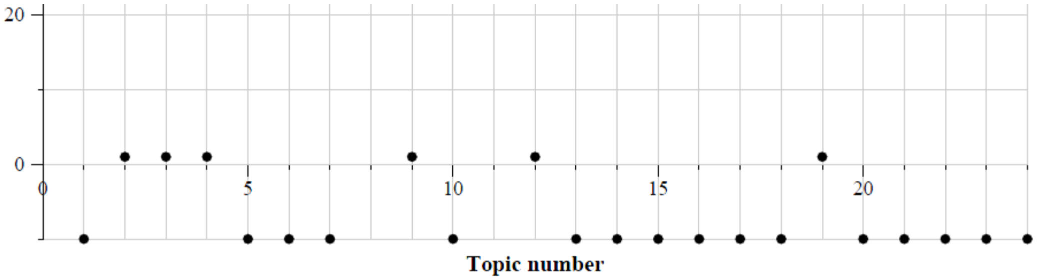


Figure 6: Graph with Average precision of first run of KIS task.

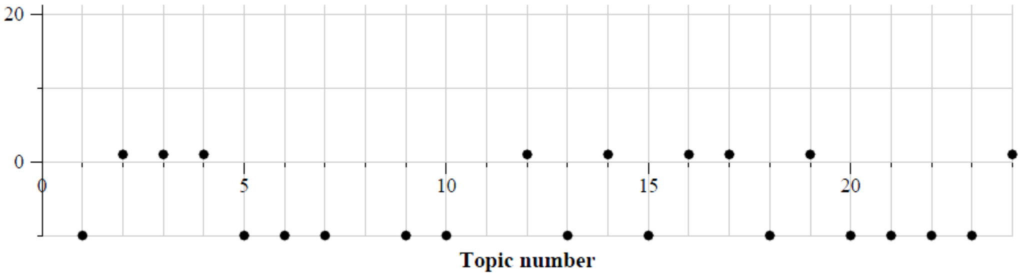


Figure 7: Graph with Average precision of second run of KIS task.

By comparing the values of Table 6, we can draw conclusions regarding the effectiveness of each of the aforementioned modules. First, by comparing runs 1 and 2 with 3 and 4, it is obvious that the use of the metadata module boosted the performance of the system. Furthermore by comparing run 1 with run 2 and run 3 with run 4, it appears that the system has benefited from lack of the high level visual concept module. However, this is not a safe conclusion to draw since, we should also take into account the fact that the limited amount of time (5 mins) in combination with the plethora of different search options might have resulted in confusion for the user.

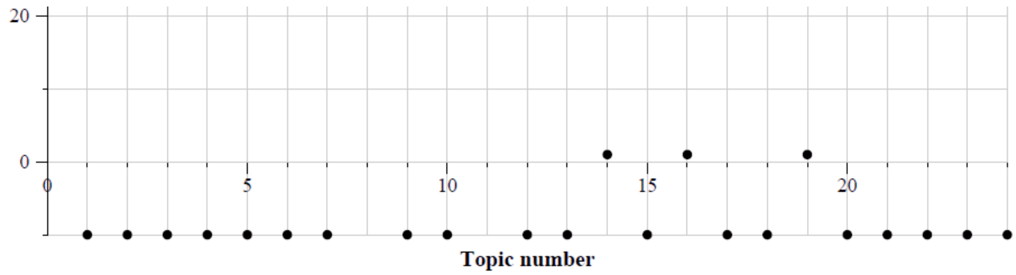


Figure 8: Graph with Average precision of third run of KIS task.

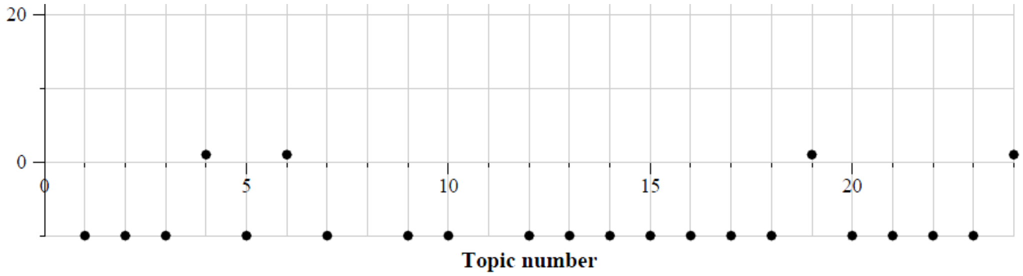


Figure 9: Graph with Average precision of forth run of KIS task.

5 Conclusions

In this paper we reported the ITI-CERTH framework for the TRECVID 2010 evaluation. ITI-CERTH participated in the SIN, KIS, INS 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 to process all video frames and not only the representative keyframes. Our submission builds upon the TRECVID 2009 HLFE task submission, tuning the previous descriptors and introducing enhancement techniques. 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 26.7% improvement of BoW+BoSW over BoW.

Regarding the TRECVID 2010 INS task, a segmentation module was integrated, in order to allow the search of specific semantic objects depicted inside the images. However, based on runs performed, we deduced that in most cases the segmentation module didn't improve the results as much as expected.

In the KIS task it is important to note that the existence of metadata played a very important role. It should be mentioned that the content-based functionalities although they provided results with common semantic or vidual characteristics with the ones requested by the query topics in most of the cases they failed to retrieve the exact video shots.

Finally, as far as the TRECVID 2010 MED task is concerned, a "model vector-based approach" has been introduced and a number of new dimensionality reduction methods have been evaluated. The proposed approach provided good recognition performance without using an elaborate and computationally expensive learning algorithm. In addition, it has been shown that the proposed dimensionality reduction methods indeed provide improved performance both in terms of recognition accuracy and system's time response.

6 Acknowledgements

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

[3] 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.

[4] 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.

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

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

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

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

[10] 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.

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

[12] 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.

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

[14] V. Mezaris, A. Dimou, and I. Kompatsiaris. On the use of feature tracks for dynamic concept detection in video. In *Proceedings of IEEE International Conference on Image Processing (ICIP 2010)*, Hong Kong, China, September 2010.

[15] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. Automatic event-based indexing of multimedia content using a joint content-event model. In *Proc. ACM Multimedia 2010, Events in MultiMedia Workshop (EiMM10)*, pages 255–258. Firenze, Italy, October 2010, 2010.

[16] 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.

[17] D. J. Kriegman P. N. Belhumeur, J. P. Hespanha. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(8):1274–1286, August 2006.

[18] M.L. Zhu and A.M. Martinez. Subclass discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(8):1274–1286, aug 2006.

[19] Xudong Jiang, Bappaditya Mandal, and Alex Kot. Eigenfeature regularization and extraction in face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):383–394, 2008.

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

[21] 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.

[22] 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.

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

[24] 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.

[25] Marijn Huijbregts, Roeland Ordelman, and Franciska de Jong. Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition. In *Proceedings of of SAMT*, Genova, Italy, December 2007.

[26] 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.

[27] 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.

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

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

[30] 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.

[31] Vasileios Mezaris, Ioannis Kompatsiaris, and Michael G. Strintzis. A framework for the efficient segmentation of large-format color images. In *In Proc. International Conference on Image Processing*, pages 761–764, 2002.

[32] Vasileios Mezaris, Ioannis Kompatsiaris, and Michael G. Strintzis. Still image segmentation tools for object-based multimedia applications. *International Journal of Pattern Recognition and Artificial Intelligence*, 18:701–725, 2004.