ITI-CERTH participation to TRECVID 2012

Anastasia Moumtzidou, Nikolaos Gkalelis, Panagiotis Sidiropoulos, Michail Dimopoulos, Spiros Nikolopoulos, Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris

Information Technologies Institute/Centre for Research and Technology Hellas, 6th Km. Xarilaou - Thermi Road, P.O. Box 60361, 57001 Thermi-Thessaloniki, Greece
{moumtzid, gkalelis, psid, midimopo, nikolopo, stefanos, bmezaris, ikom}@iti.gr

Abstract

This paper provides an overview of the tasks submitted to TRECVID 2012 by ITI-CERTH. ITI-CERTH participated in the Known-item search (KIS), in the Semantic Indexing (SIN), as well as in the Event Detection in Internet Multimedia (MED) and the Multimedia Event Recounting (MER) tasks. In the SIN task, techniques are developed, which combine video representations that express motion semantics with existing well-performing descriptors such as SIFT and Bag-of-Words for shot representation. In the MED task, two methods are evaluated, one that is based on Gaussian mixture models (GMM) and audio features, and a "semantic model vector approach" that combines a pool of subclass kernel support vector machines (KSVMs) in an ECOC framework for event detection exploiting visual information only. Furthermore, we investigate fusion strategies of the two systems in an intermediate semantic level or in score level (late fusion). In the MER task, a "model vector approach" is used to describe the semantic content of the videos, similar to the MED task, and a novel feature selection method is utilized to select the most discriminant concepts regarding the target event. Finally, the KIS search task is performed by employing VERGE, which is an interactive retrieval application combining retrieval functionalities in various modalities.

1 Introduction

This paper describes the recent work of ITI-CERTH [1] in the domain of video analysis and retrieval. Being one of the major evaluation activities in the area, TRECVID [1] has always been a target initiative for ITI-CERTH. In the past, ITI-CERTH participated in the search task under the research network COST292 (TRECVID 2006, 2007 and 2008) and in the semantic indexing (SIN) task (which is the similar to the old high-level feature extraction task) under MESH integrated project [2] (TRECVID 2008), K-SPACE project [3] (TRECVID 2007 and 2008). In 2009 ITI-CERTH has participated as stand alone organization in the HLFE and Search tasks ([2]) and in 2010 and 2011 in the KIS, INS, SIN and MED tasks ([3], [4]) of TRECVID correspondingly. Based on the acquired experience from previous submissions to TRECVID, our aim is to evaluate our algorithms and systems in order to improve and enhance them. This year, ITI-CERTH participated in four tasks: known-item search, semantic indexing, event detection in internet multimedia and multimedia event recounting tasks. In the following sections we will present in detail the applied algorithms and the evaluation for the runs we performed in the aforementioned tasks.

[1] Information Technologies Institute - Centre for Research & Technology Hellas

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

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

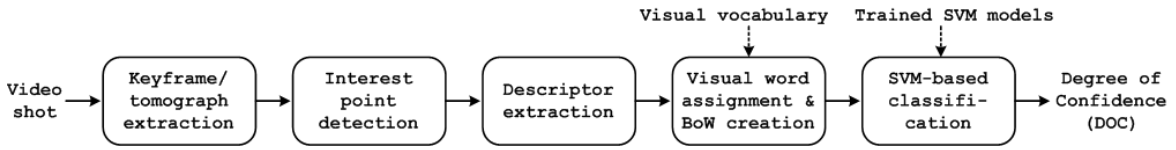

Figure 1: Block diagram of concept detection approach for one of the 25 classifiers used per shot.

2 Semantic Indexing

2.1 Objective of the submission

In TRECVID 2012, the ITI-CERTH participation in the SIN task [5] was based on an extension of our SIN 2011 technique. The main idea is to optimally combine the output of 25 linear SVM classifiers, based on multiple shot representations, descriptors, interest point detectors and assignment techniques, in order to achieve enhanced performance, both in terms of accuracy and computational cost.

This year we continued our effort in treating video as moving pictures, instead of processing only isolated key-frames (e.g. [6]). Following last year's technique ITI-CERTH used video tomographs again this year to model shot motion characteristics. Tomographs are 2-dimensional slices with one dimension in time and one dimension in space. They were used similarly to visual key-frames in distinct concept detector modules.

An important characteristic of the ITI-CERTH 2012 SIN module is the use of multiple combinations of interest point detectors, visual descriptors and assignment methods to generate several Bag-of-Words (BoW) feature vectors for each selected image of a shot (by image we mean here either a key-frame or a tomograph). For every one of these combinations a linear SVM classifier is trained, producing detection scores in the range [0,1], which express the degree of confidence (DoC) that the concept is depicted in the image. Subsequently, all DoCs for each image are combined. Furthermore, a post-processing scheme, based on the provided ontology [7], is examined. Finally, for a sub-set of 50 concepts an optimization process is introduced. For each of these concepts, the sub-set of classifiers that exhibits the best accuracy is determined and used instead of the full set of 25 developed classifiers.

2.2 Description of runs

Four SIN runs were submitted in order to evaluate the potential of the aforementioned approaches: the shot motion representation by video tomographs [8], the use of TRECVID ontology relations and the optimization process that leads to a distinct set of classifiers for each concept.

The block diagram of one classifier is given in Fig. 1. For each shot the corresponding video volume is represented either by a key-frame (as provided by TRECVID) or by a video tomograph. In the current implementation, two perpendicular tomographs are extracted, one horizontal, which tracks the visual content of a constant horizontal line through the shot frames, and one vertical, which tracks the visual content of a constant vertical line through the shot frames. An example of a vertical tomograph is given in Fig. 2. For more information about video tomographs please refer to [8]. Following the extraction of a key-frame or a tomograph, an interest point detector is employed to extract image points that will be subsequently processed. In our experiments two interest point detection variations were used. In the first variation interest points are selected through dense sampling, i.e. points were selected on a regular grid, while in the second variation interest point detection is performed through a Harris-Laplace corner detector [9]. At the resulting interest point locations, a low-level visual descriptor was calculated. In all our TRECVID 2012 runs three SIFT variations, namely SIFT, RGB-SIFT and Opponent-SIFT [10] were used. The low-level descriptors were assigned to visual words from a vocabulary that was created off-line through k-means clustering, employing both hard and soft assignment according to [11]. In all cases, the number of words for each BoW was set to 1000. A pyramidal 3×1 decomposition scheme, employing 3 equally-sized horizontal bands of the image [12], was used for every key-frame or tomograph, generating 3 different BoWs. A fourth BoW was built using the entire image. Thus, for each combination of interest point detector, descriptor and assignment method a vector of 4000 dimensions was finally extracted for each key-frame or tomograph, which is the actual input to the utilized SVM classifiers. For classification linear

Figure 2: An example of a vertical video tomograph. The frames on the left show an athlete running, followed by the camera. The vertical tomograph, shown on the right, reconstructs the shot background due to the horizontal camera motion.

SVM classifiers were employed instead of the kernel SVMs that are typically used for this task. By doing so, the computation time required for a single SVM classifier (corresponding to a single concept) decreased from 6 seconds per image (in earlier experiments with kernel SVMs) to 0.03 seconds per image, which is a significant improvement, given the volume of data that we needed to process. All classifiers were trained off-line, using the extensive training data provided by the organizers of the TRECVID SIN task. As in past participations of ours, a variable proportion of positive/negative samples was used. The proportion ranged from 1:5 to 1:1. However, the maximum limit of 20000 positive and negative vectors for each concept, which we adopted in our 2011 SIN participation due to computational limitations, was lifted in our 2012 experiments. The output of the classifier is a Degree of Confidence (DoC) score for the corresponding concept, which expresses the estimated probability that the concept is present in the current shot.

The above process is iterated for some or all of the 25 classifiers, depending on the run (see Table 1). A distinct local-feature-based classifier is constructed based on the representation, interest point detection, low-level descriptor and assignment strategy combinations of Table 1. A 25th classifier that uses global visual descriptors (HSV histograms) was also employed. The overall Degree of Confidence for each concept is estimated as the harmonic mean of the individual classifiers' DoCs. Final results per high-level feature were sorted by DoC in descending order and the top 2000 shots per concept were submitted to NIST.

Table 1: The 25 employed classifiers.

Representation	**Classifiers**
Key-frame	12 local-image-feature-based Classifiers (3 descriptors (SIFT, Opponent-SIFT, RGB-SIFT) x 2 sampling strategies (Dense, Harris-Laplace) x 2 BoW strategies (soft-, hard-assignment)) 1 global-image-feature-based classifier (color histograms)
Tomographs	12 tomograph-based Classifiers (2 types of video tomographs (horizontal, vertical) x 3 descriptors (SIFT, Opponent-SIFT, RGB-SIFT) x 2 BoW strategies (soft-, hard-assignment))

In two of the submitted runs, the use of the provided ontology was also tested. Regarding "imply" relations our methodology was the same as last year [4]. Furthermore, in this year's SIN task we also employed some of the "exclude" ontology relations. Specifically, the top 5000 shots of "Junk Frame" concept (as returned by our "Junk Frame" detector) were discarded from further consideration when analyzing all concepts that "Junk Frame" excludes.

Finally, one run included an optimization technique for choosing the set of classifiers used for each concept. More specifically, using a genetic algorithm optimization process, for each of the 50 concepts that were evaluated in TRECVID 2011 SIN task we estimated the sub-set of the 25 classifiers that achieves the best detection accuracy. We then used this set of classifiers for generating the corresponding results in TRECVID 2012, instead of using the full set of 25 classifiers for each of these concepts.

The 4 submitted runs for the full submission and the two runs submitted for pair submission were:

- ITI-CERTH-Run 1: "Optimized run; Optimized combination of up to 25 classifiers per concept, for 50 of the concepts, and partial use of concept ontology". In this run all 25 classifiers of Table 1 were used. The averaging of the 25 confidence scores, which are retrieved for each pair of shot-concept, is conducted through the estimation of their geometric mean. The results are additionally filtered using

the aforementioned ontology-based technique. The 50 concepts that were evaluated in TRECVID 2011 SIN Task were optimized using the genetic algorithm described previously (thus each concept employing an optimized set of up to 25 of the available classifiers).

- ITI-CERTH-Run 2: "Improved run; Combination of 25 classifiers per concept, and partial use of concept ontology". This run is similar to run 1, the only difference being that the optimization step is omitted. Thus, exactly 25 classifiers per concept are used for all 346 concepts.
- ITI-CERTH-Run 3: "Extended run; Combination of 25 classifiers per concept". This run is similar to run 2, the only difference being that the final ontology-based post-processing step is omitted.
- ITI-CERTH-Run 4: "Baseline run; Combination of 13 classifiers per concept." This is a baseline run using only visual key-frames. The 12 key-frame based classifiers of Table 1 and the one based on low-level visual descriptors were employed.
- ITI-CERTH-Run 7: "Concept pair optimized run; Product Rule". This is a run that employs the results of run 1 of the single concept task and the product rule to generate the pair results.
- ITI-CERTH-Run 8: "Concept pair baseline run; Product Rule". This is a run that employs the results of run 4 of the single concept task and the product rule to generate the pair results.

2.3 Results

The runs described above were submitted for the 2012 TRECVID SIN competition. The evaluation results of the aforementioned runs are given in terms of the Mean Extended Inferred Average Precision (MXinfAP) both per run and per high level feature. Table 2 summarizes the results for each run presenting the MXinfAP of all runs. The main drawback of using linear SVMs is related to the

Table 2: Mean Extended Inferred Average Precision (MXinfAP) for all high level features and runs.

	ITI-CERTH 1	**ITI-CERTH 2**	**ITI-CERTH 3**	**ITI-CERTH 4**
MxinfAP	**0.132 (0.162)**	**0.131 (0.161)**	**0.115 (0.156)**	**0.135**
MxinfAP Light	**0.153 (0.18)**	**0.148 (0.174)**	**0.131 (0.171)**	**0.162**
	ITI-CERTH 7		**ITI-CERTH 8**	
MxinfAP Pair	**0.018**		**0.014**	

simplicity of the boundary hyper-surface that separates the two classes, which in this case is a hyper-plane. Furthermore, images with simple texture and usually with no semantic content tend to generate clusters of similar features that, when summed up in histograms, fall to a small number of bins. Thus, the resulting feature vectors are biased towards some dimensions, which accumulate the majority of the feature vector power. As a result, the use of a simple hyper-plane to assign a probability value as a function of the distance between the feature vector and the hyper-plane is expected to lead to values that would be extremely biased, either towards 0 or towards 1. In the second case, a false positive that would be ranked in top positions when sorting the results is generated.

We have developed two approaches to overcome this shortcoming. The first one is to use the probabilities of "Junk Frame" concept to discard the frames that are most likely to be junk (thus erroneous). The second is to discard shots of small duration, based on the fact that these frames are often generated by over-segmentation mistakes in the shot segmentation process. Following our preliminary experiments in the TRECVID 2011 setup, we chose to follow the second approach only in our 2012 "baseline run" (ITI-CERTH 4). The 2012 results, in which this "baseline run" demonstrated the best performance among our 4 runs, revealed that this was not the optimal decision and further analysis of this drawback was required. In Table 2, inside parentheses, we report the MXinfAP performance of each run in the full and light task, having discarded the shots of small duration. These scores justify the contradictory evaluation results that found "baseline run" as the most accurate configuration. If in all 4 runs the same strategy was followed, the total accuracy gain of run "ITI-CERTH 1" in comparison to "ITI-CERTH 4" would have been 20%. Similarly, just introducing the tomographs in our baseline (i.e. comparing the performance difference between "ITI-CERTH 3" and "ITI-CERTH 4" runs) would result in a 15.5% accuracy boost. These gains are in line with our assumption that the spatio-temporal slices can express the shot motion patterns in a meaningful way. Furthermore, the use of the ontology further enhances the performance by 3.2%, while the use of

the optimization seems to only increase it marginally. This can be explained by the fact that the optimization was followed for only 50 of the 346 concepts, out of which only 7 were included in the set of the evaluated concepts.

Finally, the 0.162 MXinfAP score would bring CERTH in the 9th place, among 15 participants in the full run track. This median performance is judged as satisfactory, taking into account that we made design choices that favor speed of execution over accuracy, such as the use of linear SVMs.

3 Multimedia Event Detection

3.1 Objective of the submission

High level event detection in video is a challenging task that has practical applications in several domains such as news analysis, video surveillance and multimedia organization and retrieval. Applications in those domains are often time-critical and the use of systems that provide low-latency responses are highly desired. The objective of our submission is the evaluation of a number of efficient algorithms that we have recently developed for the task of event detection.

3.2 System Overview

The target of the event detection system is to learn a decision function $f(\mathbf{X}) \rightarrow \mathcal{Y}$, $\mathcal{Y} = \{1, 2\}$ that assigns the test video $\mathbf{X}$ to an event class (labelled with the integer 1) or to the "rest of the world" class (labelled with the integer 2).

3.2.1 Subsystem exploiting visual information

A concept-based approach is used to represent a video with a sequence of model vectors similarly to [13, 14, 4]. Our method exploits only static visual information extracted from selected video frames following the procedure described in section 2. To be more precise, first we decode the video signal and select one frame every 6 seconds with uniform time sampling in order to represent the video with a sequence of keyframes. Then, a dense sampling strategy is combined with a spatial pyramid approach to extract salient image points at different pyramid levels, the opponentSIFT color descriptor is used to derive 384-dimensional feature vectors at those points, and a bag-of-words (BoW) method is applied to construct a visual codebook. Subsequently, a video frame is described with a feature vector in $\mathbb{R}^{4000}$ using soft assignment of visual words to image features at each pyramid level. The final feature vectors are used as input to a set $\mathcal{G} = \{h_\kappa()|\kappa = 1, \ldots, F\}$ of $F = 172$ trained semantic concept detectors $h_\kappa() \rightarrow [0, 1]$ for associating each keyframe with a model vector [13]. In particular, we used a subset of the concept detectors derived from our participation in the SIN task (section 2) referring to the concepts depicted in Table 3. Following the above procedure, the p-th video of the

Table 3: The 172 concepts of the SIN 2012 task used by the event detection systems.

3 Or More People, Adult Male Human, Adult Female Human, Airplane, Animal, Apartments, Armed Person, Athlete, Attached Body Parts, Baby, Beach, Beards, Bicycles, Bicycling, Birds, Boat Ship, Car, Celebrity Entertainment, Chair, Charts, Cheering, Child, City, Cityscape, Classroom, Clearing, Clouds, Commercial Advertisement, Computer Or Television Screens, Computers, Conference Room, Construction Vehicles, Construction Worker, Crowd, Dancing, Daytime Outdoor, Desert, Dogs, Door Opening, Doorway, Dresses, Driver, Eaters, Exiting A Vehicle, Exiting Car, Explosion Fire, Face, Female Human Face Closeup, Female Human Face, Female Person, Female Reporter, Fields, Flags, Flowers, Food, Forest, Free Standing Structures, Furniture, Girl, Glasses, Graphic, Greeting, Ground Combat, Ground Vehicles, Gun, Gym, Hand, Harbors, Head And Shoulder, Highway, Hill, Hockey, Human Young Adult, Indoor, Indoor Sports Venue, Infants, Joy, Junk Frame, Kitchen, Laboratory, Lakes, Landscape, Legs, Male Human Face Closeup, Male Human Face, Male Person, Male Reporter, Man Made Thing, Man Wearing A Suit, Maps, Meeting, Military, Military Base, Military Personnel, Minivan, Motorcycle, Mountain, News Studio, Nighttime, Oceans, Office, Old People, Outdoor, People Marching, Person Drops An Object, Plant, Police Private Security Personnel, Politicians, Press Conference, Religious Building, Religious Figures, Reporters, Residential Buildings, Rifles, Road, Road Overpass, Roadway Junction, Rocky Ground, Room, Running, Sadness, School, Science Technology, Scientists, Shopping Mall, Singing, Single Person Female, Single Person Male, Sitting Down, Skating, Ski, Sky, Snow, Soccer, Soccer Player, Sofa, Speaker At Podium, Speaking To Camera, Sports, Sports Car, Stadium, Standing, Streets, Suburban, Suits, Sunglasses, Sunny, Swimming, Table, Talking, Teenagers, Telephones, Tent, Text, Throwing, Tower, Traffic, Two People, Underwater, Urban Park, Valleys, Van, Vehicle, Walking, Walking Running, Waterscape Waterfront, Weapons, Wild Animal, Windows.

i-th class, consisting of o_p keyframes, is represented as $\mathbf{X}_i^p = [\mathbf{x}_i^{p,1}, \ldots, \mathbf{x}_i^{p,o_p}]$, where $\mathbf{x}_i^{p,q} \in \mathbb{R}^F$ is the model vector of the q-th keyframe of the video. As a last step, we retrieve the final feature vector $\mathbf{x}_i^p$ for representing the whole video by averaging the model vectors along all keyframes $\mathbf{x}_i^p = \sum_{q=1}^{o_p} \mathbf{x}_i^{p,q}$; note that the κ-th element of the feature vector of the video expresses the degree of confidence that the κ-th semantic concept is depicted in the video.

The event categories are learned using an approach that splits the event class to several subclasses and embeds a pool of subclass kernel support vector machines (KSVM) [15] in the error correcting output code framework (ECOC) [16, 17, 18] as explained in the following. First, we exploit an iterative algorithm to derive a subclass division of the training data that belong to the event class [19, 18]. Starting from the initial target event class partition at each iteration a new partition $\mathcal{X}^{(r)} = \{\mathcal{X}_j \| j = 1, \ldots, r\}$ is created by increasing the number of subclasses of the event class by one, where $r = 1, \ldots, R$ is the iteration index, and R is the total number of iterations. At each iteration the following nongaussianity measure is computed

$$\Phi^{(r)} = \frac{1}{C} \sum_{i=1}^{C} (\gamma_j^{(3)} + \gamma_j^{(4)}) \tag{1}$$

The quantities $\gamma_j^{(3)}$ and $\gamma_j^{(4)}$ are estimates of the multivariate standardized skewness and kurtosis of the j-th subclass of the event, computed as follows $\gamma_j^{(3)} = \frac{1}{F} \sum_{\kappa=1}^{F} |\gamma_{j,\kappa}^{(3)}|$, $\gamma_j^{(4)} = \frac{1}{F} \sum_{\kappa=1}^{F} |\gamma_{j,\kappa}^{(3)} - 3|$, where, $\gamma_{j,\kappa}^{(n)} = (\frac{1}{N_j} \sum_{x_\kappa \in \mathcal{X}_j} (x_\kappa - \mu_{j,\kappa})^n) / \sigma_{j,\kappa}^n$ are estimates of the skewness ($n = 3$) and kurtosis ($n = 4$) of the j-th subclass along the κ-th dimension. In the above equation x_κ is the κ-th element of the video model vector $\mathbf{x}$, and $\mu_{i,j,\kappa}, \sigma_{i,j,\kappa}$, N_j are the sample mean, standard deviation and number of video model vectors of j-th subclass along the κ-th dimension. At the end of this iterative algorithm, the best partition $\mathcal{X}^{(H_1)}$ is selected according to the following rule

$$\mathcal{X}^{(H_1)} = \underset{r}{\operatorname{argmin}} (\Phi^{(r)}) \,, \tag{2}$$

where, H_1 is the optimal number of subclasses of the event class.

Next, we train one KSVM with radial basis function (RBF) kernel [15] for each subclass of the event class, i.e. the j-th KSVM is trained using as positive samples the video model vectors of the j-th subclass and as negative samples the video model vectors of the "rest of the world" class. For the KSVMs we used the libsvm library [20], in particular we exploit the implementation that provides a probability estimate in the output of the KSVM. The derived set of dichotomizers, $g_j \in [0, 1], j = 1, \ldots, H_1$ is then combined using a subclass error correcting output code (SECOC) framework [16]. In particular, an one-versus-one SECOC design is applied exploiting the loss-weighted decoding measure as explained in the following. During the coding process, each subclass is associated with a row of the ternary coding matrix $\mathbf{M} \in \{1, 0, -1\}^{H \times H_1}$, where $H = H_1 + 1$ is the total number of subclasses. During the decoding stage a test video model vector $\mathbf{z}^\tau$ is classified to one of the subclasses by first evaluating the H_1 dichotomizers in order to create a codeword for it, and then comparing the derived codeword with the base codewords in the coding matrix. For the comparison of the codewords we use the loss-weighted decoding measure [17]:

$$d_j^\tau = \sum_{s=1}^{H_1} M_j^s g_s(\mathbf{z}^\tau) \tilde{M}_j^s, \;\; j = 1, \ldots, H_1 \,, \tag{3}$$

where $M_j^s, \tilde{M}_j^s$ are the elements of the coding and weighting matrix respectively, that correspond to the j-th subclass and the dichotomizer that separates the s-th subclass from the "rest of the world" class. To this end, we derive a confidence score regarding the presence of the event in the test video by averaging the derived similarity values d_j^τ along the event subclasses $d^\tau = \frac{1}{H_1} \sum_{j=1}^{H_1} d_j^\tau$. We should note that $M_j^s \in \{0, 1\}$, $\tilde{M}_j^s \in [0, 1]$ for $j = 1, \ldots, H_1$, $s = 1, \ldots, H_1$, and, therefore $d_j^\tau \in [0, 1]$. The test video shot is then classified to the target event according to the rule $d^\tau \geq \theta$, where $\theta \in [0, 1]$ is the detection threshold value.

3.2.2 Subsystem exploiting audio information

In parallel to the technique of section 3.2.1, a method that exploits only low-level audio information extracted from videos is used to perform event detection. This method [21] is based on short-time frequency analysis of audio using linear frequency cepstral coefficients (LFCC) and modulation spectrogram (MSG). These features are complementary to each other in the sense that LFCCs are calculated for a very short time window and MSGs for a longer one.

First, for each audio frame we extract 20 static LFCCs, then compute their first (delta) and second order derivative (double delta) coefficients, leading to a 60-element feature vector. For MSG we compute a short-time FFT spectrogram, divide the frequency range in 18 (sub-)bands and in each band apply two filters, a low-pass and a band-pass. As a result we have 2 coefficients for each band leading in a 36-element feature vector for every audio frame. Next, we proceed with a normalization scheme in a file-by-file manner. For every coefficient c_i^f of every frame f of the audio file we subtract the mean μ_i and divide by the standard deviation σ_i that corresponds to the coefficient c_i. Mean and standard deviation are computed among all frames of a file for every coefficient c_i. Normalization is applied separately for LFCC and MSG. Then, we sample the frames that belong to each file according to an energy detection process. A threshold is determined in an adaptive manner and we only keep those frames whose energy is adequate to be characterised non-silent.

Training is based on Gaussian Mixture Models (GMM) and consists of two phases [22]. First, we train a so-called Universal Background Model GMM (UBM-GMM) using positive examples of all relevant events E06-E15 and E21-E30, as well as an equal number of negative examples (videos that do not belong to any event). In the second step, the training set is used to train a GMM model per event via Maximum a Posteriori (MAP) adaptation from the UBM-GMM. The above training process is repeated separately for MSG and LFCC features (thus resulting in 2 sets of GMMs).

In the testing phase we extract the same features as the ones for training, apply normalization and sampling with energy detection for all test videos. Test video feature vectors are then used to derive a log-likelihood ratio (LLR) score using the event GMMs [21]. LLR score values are in range $(-\infty, +\infty)$. Careful examination of LLR score results shows that the number of files with LLR below -1 or above +1 is very low. So, instead of scaling all test video LLR scores to $[0, 1]$ based on the overall maximum and minimum values, we decided to floor and ceil the values that are below -1 and above $+1$ respectively, and scale the resulting values to the range $[0, 1]$ to retrieve a confidence score for the test videos. As a final step, we combine the scores of the two GMMs referring to the same event using a weighted average strategy.

3.3 Dataset description

The following 5 video collections of the TRECVID MED 2012 track are used for evaluating the 20 event detection systems for the Pre-Specified event task:

- EVENTS: This collection contains approximately 200 videos for each Pre-Specified event[1]. It is the union of the MED11 Training event collection (event kits referring to events E06-15) and the MED12 Pre-Specified events (event kits referring to events E21-E30).
- OTHER-EVENTS: The MED11 Training event collection containing approximately 820 videos belonging to one of the 5 training events E01-05.
- DEV-T: The MED11 transparent data collection (DEV-T) contains 10273 videos ($\sim$ 350 hours) belonging to one of the events E01-05 or to the "rest of the world" category.
- MED11TEST: The MED11 Test collection contains 32061 videos ($\sim$ 1000 hours) belonging to one of the events E06-15 or to the "rest of the world" category.
- PROGTEST: The MED12 Progress Test collection consists of 98118 videos ($\sim$ 4000 hours) belonging to one of the events E06-15, E21-E30 or to the "rest of the world" category.

[1]The Pre-Specified events are: E06: Birthday party, E07: Changing a vehicle tire, E08: Flash mob gathering, E09: Getting a vehicle unstuck, E10: Grooming an animal, E11: Making a sandwich, E12: Parade, E13: Parkour, E14: Repairing an appliance, E15: Working on a sewing project, E21: Attempting a bike trick, E22: Cleaning an appliance, E23: Dog show, E24: Giving directions to a location, E25: Marriage proposal, E26: Renovating a home, E27: Rock climbing, E28: Town hall meeting, E29: Winning a race without a vehicle, E30: Working on a metal crafts project

The first 4 sets described above (EVENTS, OTHER-EVENTS, DEVT, MED11TEST) are designated for training purposes, while the last set (PROGTEST) is used for the blind evaluation of the algorithms. The ground truth annotation tags used for declaring the relation of a video clip to a target event are "positive", which denotes that the clip contains at least one instance of the event, "near miss", to denote that the clip is closely related to the event but it lacks critical evidence for a human to declare that the event occurred, "related" to declare that the clip contains one or more elements of the event but does not meet the requirements to be a positive event instance and "not sure". In case that the clip is not related with any of the target events the label "NULL" is used. During the training procedure, we treated the clips that are annotated as "near miss", "related", or "not sure" as positive instances of the event. We should also note that the training video collection is a very unbalanced set, i.e. the number of negative instances greatly outnumbers the number of positive instances for each event.

3.4 Description of runs

We submitted 4 runs in the TRECVID MED 2012 evaluation track, namely, p-visual, c-audio, c-audiovisual and c-audiovisualLate. Our primary run (p-visual) utilizes a pool of semantic concept detectors trained using only static visual information (section 3.2.1), while our second run exploits audio information alone (section 3.2.2). The third (c-audiovisual) and fourth run (c-audiovisualLate) perform fusion of the audio and visual information in an intermediate and score level respectively.

- p-visual: In this run we apply the method described in section 3.2.1. That is, the semantic concept detectors exploiting static visual information are used to describe the semantic content of the video, and the detection of an event is done using a pool of RBF KSVMs under the ECOC framework. The parameters of the detection algorithm are learned from grid search using a 3-fold cross validation procedure, where overall optimization procedure is guided by the Normalized Detection Cost (NDC). For this purpose we use the 4 sets dedicated for training purposes described in section 3.3 (EVENTS, OTHER-EVENTS, DEVT, MED11TEST), where at each fold, for each event, we randomly split the overall set to 50% training set, 20% validation set and 30% test set. Then, we use the overall set to learn the final parameters of the detection algorithm for each event and apply the detection algorithm to the MED12 PROGTEST.

- c-audio: The method described in section 3.2.2 is used in this run. In particular, we use the EVENTS set, the positive examples of MED11TEST and a randomly selected set of negative examples from the MED11TEST to build the UBM-GMM, and then adapt one GMM for each event using the EVENTS set. The parameters of the system are learned using a cross-validation procedure with 70% training and 30% testing random splits. The trained event GMMs are applied in the PROGTEST dataset and the LLR scores are transformed to confidence scores as explained in section 3.2.2.

- c-audiovisual: In this run we exploit the audiovisual video information by combining the two runs described above (p-visual, c-audio) in an intermediate level. In particular, the video model vectors (section 3.2.1) are extended using the event confidence scores that are derived using audio information (section 3.2.2), to yield a new feature vector of the video. The new video feature vectors are then used to learn and detect the events exploiting the approach described in section 3.2.1. The parameters of the detection algorithm are learned following a training procedure similar to the one used in our primary run (p-visual).

- c-audiovisualLate: In contrary to the above run, here we combine visual and audio information in the score level (late fusion). In particular, the confidence scores derived from the two former runs described above (p-visual, c-audio) are combined by averaging them to yield the final detection score for the test video.

3.5 Results

The evaluation results of our 4 runs in the TRECVID MED 2012 Pre-Specified event task are shown in Table 4, in terms of NDC, probability of false alarms P_{FA} and probability of missed detections P_{MD},

averaged along the 20 Pre-Specified events. Moreover, in the last column we provide the number of events (out of the 20 defined events) for which we reached the goal of the evaluation, i.e. to achieve error rates for P_{MD} and P_{FA} values below 4% and 50% respectively. From the analysis of the detection

Table 4: Evaluation results.

Run	*NDC*	P_{FA}	P_{MD}	*# Events*
p-visual	0.9088	0.0009	0.8980	8
c-audio	1.8637	0.0856	0.7953	1
c-audiovisual	0.9013	0.0007	0.8929	10
c-audiovisualLate	0.9222	0.0009	0.9113	3

results we can conclude the following:

- The overall performance of our runs compared to the rest of the submissions is rather average. However, taking into account that in comparison to the other submissions we use only limited visual features (opponentSIFT descriptors in sparsely sampled keyframes), the performance of our detection algorithms can be considered good.
- Among the 10 runs that exploit only static visual information our respective run (p-visual) ranks 4th (Fig. 3). Moreover, in this run we exploit only one descriptor (opponentSIFT), while in most of the other runs more than one static visual descriptors are used.
- The run using only audio information (c-audio) was ranked in the lower quartile of the submissions. It is the only run submitted to the competition that uses exclusively audio information. On a mean NDC basis, it outperforms a few runs that are combinations of audio, visual, video and text modalities. Optimization of the threshold values per event was done according to the 12.5:1 miss-to-false alarm ratio without taking NDC directly into consideration. Analysis of the statistics for the MED'11 task with various thresholds done after the submission showed that if threshold optimization were done on an NDC basis the run could have scored just above 1.
- We observe that in the run where we combine visual and audio information at an intermediate level of the detection process, we get improved performance compared to runs that exploit visual or audio information alone. However, this is not the case for the run that combines the two modalities in the score level (late fusion). We suspect that the degrading in performance in that case is due to the use of simple averaging (e.g. as opposed to using weighted average aggregation).
- The event agent execution time of our algorithms for processing the whole PROGTEST dataset for one event is in the order of a few minutes. Therefore, we conclude that our event detection system offers real-time performance. Moreover, our run that exploits only audio information (c-audio) is among the fastest of the submitted runs (according to the total processing times reported in the submitted runs).

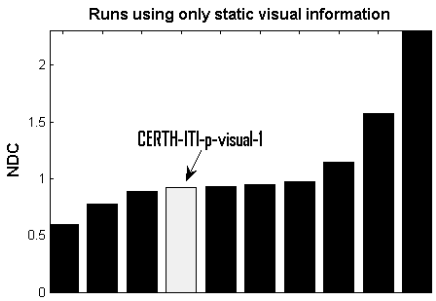

Figure 3: Actual NDC for all submissions that use only static visual information.

4 Multimedia Event Recounting

4.1 Objective of the submission

Objective of our submission is the evaluation of our recently developed algorithm for the task of event recounting. This is an extension of event detection techniques developed in our group [23, 4].

4.2 System Overview

The goal of the event recounting system presented in this section can be stated as follows: given a pool of F semantic concept detectors, $\mathcal{G} = \{h_\kappa() | \kappa = 1, \ldots, F\}$, and an annotated video collection $\mathcal{X} = \{\mathbf{X}_i^p | p = 1, \ldots, L_i,\ i = 1, 2\}$ belonging to the target event ($i = 1$) or to the "rest of the world" class ($i = 2$), where $\mathbf{X}_i^p$ denotes the p-th video of the i-th class and L_i is the number of videos belonging to the i-th class, annotate an unlabelled video signal $\mathbf{X}$ containing the target event with a small fraction of $I < F$ semantic concepts related with the event. To this end, we exploit the concept-based approach described in section 3.2.1 to represent a video signal $\mathbf{X}_i^p$ with a video model vector $\mathbf{x}_i^p \in \mathbb{R}^F$, where $F = 172$ is the set cardinality of our SIN concept detectors selected for our participation in the MED Task (Table 3). We then utilize the nongaussianity criterion and a mixture subclass discriminant analysis (MSDA) algorithm [19, 24], to acquire a subclass division of the data of $H = 3$ total subclasses, and derive a transformation matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2] \in \mathbb{R}^{F \times H-1}$. Inspired from [25], the derived transformation matrix is used for computing the most discriminant concepts concerning the target event depicted in a test video as explained in the following. First we compute the weighted video model vector $\mathbf{y}^\tau = [y_1^\tau, \ldots, y_F^\tau]^T$ using

$$\mathbf{y}^\tau = \text{argmax}(\mathbf{w}_1 \circ \mathbf{x}^\tau, \mathbf{w}_2 \circ \mathbf{x}^\tau) \tag{4}$$

where y_f^τ express the degree of confidence (DoC) concerning the f-th concept weighted with a significance value regarding the target event, $\mathbf{x}^\tau$ is the video model vector, and the operator $\circ$ is used to denote element-wise vector multiplication. The I most discriminant concepts are then selected according to the following rule

$$\{c_1, \ldots, c_I\} = \text{argmax}_I(y_1^\tau, \ldots, y_F^\tau)\,. \tag{5}$$

In the MER evaluation we fixed the number of discriminant concepts for describing a test video to $I = 15$.

4.3 Dataset description

For the MER evaluation task two video collections are defined:

- MER Evaluation Test Set: This set consists of 30 videos, i.e. 6 videos from each of the 5 MER events, namely E22, E26, E27, E28 and E30.
- MER Progress Test Set: This set consists of all videos in the MED PROGTEST collection for which our primary run MED system (p-visual, section 3.4) identified them as positive event clips, for all of the five MER events.

4.4 Description of the run

A single run was submitted in the TRECVID MER 2012 evaluation track. The method described in section 4.2 is used to provide an event recounting for each video and each event in the MER Evaluation Test Set and MER Progress Test Set. The output files are generated in the required format using a perl script. Specifically, this script receives as input a text file containing the names and ids of the detected concepts as well as the corresponding DoCs and generates as output an XML file according to the DTD schema provided by NIST.

4.5 Results

Two metrics are used for evaluating the MER runs: a) MER-to-event: fraction of MER outputs for which the judges identified correctly the underlying MER event, b) MER-to-clip: fraction of MER outputs for which the judges identified correctly the corresponding clip. The evaluation results of our run for each event are depicted in Table 5. From the attained results we conclude that the performance

Table 5: Evaluation results.

Event	*MER-to-event*	*MER-to-clip*
E22	18.1818%	33.3333%
E26	12.8205%	48.7179%
E27	72.7273%	15.1515%
E28	100%	17.9487%
E30	3.0303%	9.0909%
All	42.3729%	25.4237%

of our run is rather average. This was expected for two reasons: a) only a small set of concept detectors is used to describe the semantic content of the videos, b) our concept detectors exploit only limited static visual information; therefore, noisy detections are derived. On the other hand, for some events (E27, E28) very promising results were attained regarding the MER-to-event ratio. Considering the above analysis further investigation of the proposed approach seems to be worthwhile.

5 Known Item Interactive Search

5.1 Objective of the submission

ITI-CERTH's participation in the TRECVID 2012 known-item (KIS) task aimed at studying and drawing conclusions regarding the effectiveness of different ways of interface representation and retrieval modules, which are integrated in an interactive video search engine, in the retrieval procedure. Shot-based video retrieval has been the standard representation for most of the video retrieval systems participating in TRECVID. This was also dictated by the past search tasks, in which the users were asked to retrieve shots that satisfy specific criteria. In the recent years, the KIS task introduced by TRECVID has changed this requirement and expects retrieval at video level. Motivated by this we compared shot and video-based representations and considering a variety of retrieval modalities. Within the context of this effort, four runs were submitted, each combining existing modules and presentation formats in a different way, for evaluation purposes.

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

5.2 System Overview

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

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

- High Level Concept Retrieval;

The search system is built on open source web technologies, more specifically Apache server, PHP, JavaScript, mySQL database, Strawberry Perl and the Indri Search Engine that is part of the Lemur Toolkit [26].

Besides the basic retrieval modules, VERGE integrates a set of complementary functionalities, which aim at improving the retrieved results. To begin with, the system supports two types of data representation: shot and video-based. In the shot-based representation, VERGE presents all video shots in its main window, while the search modules (except from metadata search, which are video-based by default), provide results at shot level. On the other hand, when the video representation is selected, each video is visualized by its middle shot (i.e. represented the middle keyframe of this shot) and a preview by rolling at most four of its shots. In this case only the video-based retrieval techniques are considered (i.e. Transcription module, Metadata module, Video Indexing based on ASR and metadata and High Level Visual Concepts module).

Moreover, regardless of the type of representation, the system supports basic temporal queries, such as the shot-segmented view of each video. Also, the shots/videos selected by a user are stored in a storage structure that mimics the functionality of the shopping cart. Finally, a history bin is supported, in which all the user actions are recorded.

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

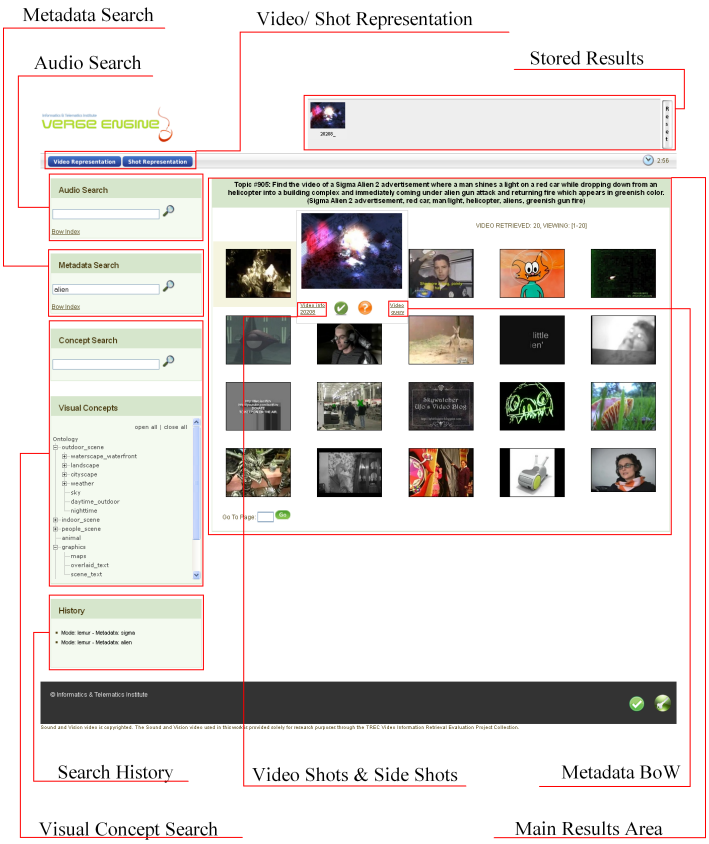

Figure 4: User interface of the interactive search platform using video representation.

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

5.2.1 Visual Similarity Search Module

The visual similarity search module performs image content-based retrieval with a view to retrieving visually similar results. Following the visual similarity module implementation in [3], we have chosen an MPEG-7 schemes that relies on color and texture (i.e. ColorLayout and EdgeHistogram were concatenated). It should be noted that this module is only available when shot representation is selected.

5.2.2 Transcription Search Module

The textual query module exploits the shot audio information. To begin with, Automatic Speech Recognition (ASR) is applied on test video data. In this implementation, the ASR is provided by [27]. The textual information generated is used to create a full-text index utilizing Lemur [26], a toolkit designed to facilitate research in language modeling. This module is available both in the shot and video representation. When the video representation is selected, the video associated text consists of the concatenation of the transcriptions for all te included shots and is used for creating a full-text video index.

5.2.3 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 collection, a metadata file is provided that contains a title, a subject, a set of keywords and a short description provided by the owner, which are usually relevant to the content of the video. The first step of metadata processing involves the extraction of this content from the files with parsing. The next step deals with the processing of the acquired content and includes punctuation and stop words removal. Finally, the processed content is indexed with the Lemur toolkit that enables fast retrieval as well easy formulation of complicated queries in the same way described in section 5.2.2.

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

For implementing the "Video Query" functionality, we have employed a bag-of-words (BoW) representation of video metadata. More specifically, in order to express each video as a bag-of-words we initially pre-processed the full set of metadata for removing stop words and words that are not recognized by WordNet [28]. Then, by selecting the 1000 most frequent words to define a Codebook of representative words, we have expressed each video as an occurrence count histogram of the representative words in its metadata. Finally, probabilistic Latent Semantic Analysis [29] was applied on the semantically enhanced video representations to discover their hidden relations. The result of pLSA was to express each video as a mixture of 30 latent topics, suitable for performing indexing and retrieval on the full video collection. For indexing new video descriptions, such as the as the ones provided by the user in the "Transcription Search Module", we have followed the pLSA theory that proposes to repeat the Expectation Maximization (EM) steps [30] that have been used during the training phase, but without updating the values of the word-topic probability distribution matrix.

5.2.5 High Level Visual Concept Retrieval

This module facilitates search by indexing the video shots based on high level visual concept information, such as water, aircraft, landscape and crowd. Specifically, we have incorporated into the system all the 346 concepts studied in the TRECVID 2012 SIN task using the techniques and the algorithms described in detail in section 2. It should be noted that in order to expand the initial set of concepts, we inserted manually synonyms that could describe the initial entries equally well (e.g. as synonyms of the concept "demonstration" were considered "protest" and "riot").

This module by default is applied at shot level. However, in order to extend its use at video level as well, we fused the scores provided at shot level and generated concept scores for each video. Specifically, based on the assumption that the important information for a video is whether a concept exists or not (and not how many times it appears) we simply assigned the greater confidence value between the shots of one video for a certain concept.

5.3 Known-Item Search Task Results

The system developed for the known-item search task includes all the aforementioned modules. We submitted four runs to the Known-Item Search task. These runs employed different combinations of the existing modules and are described in Table 6. It should be mentioned that the complementary

Table 6: Modules incorporated in each run.

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

functionalities (i.e. shot-segmented view of each video, storage structure and history bin) were available in all runs. According to the TRECVID guidelines the time duration for each run was set to five minutes. The number of topics and the mean inverted rank for each run are illustrated in Table 7. By

Table 7: Evaluation of search task results.

Run IDs	Mean Inverted Rank	Number of correctly recognized topics
I_A_YES_ITI-CERTH_1	0.542	13/24
I_A_YES_ITI-CERTH_2	0.417	10/24
I_A_YES_ITI-CERTH_3	0.667	16/24
I_A_YES_ITI-CERTH_4	0.500	12/24

comparing the values of Table 7, we can draw conclusions regarding the effectiveness of each of the aforementioned modules. First, by comparing runs 1 and 3 with 2 and 4, it is obvious that video-based representation runs seems to achieve higher performance compared to the shot-based runs. A possible explanation might be that video representation allows the easier and faster preview of videos (through rolling) and the video collection in total. The difference between runs 1 and 3 is the existence or not of BoW module in both metadata and ASR text. The same applies to run 2, which differs from run 4 in the existence of the high level visual concept module. In both cases 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.

Compared to the other systems participated in interactive Known Item Search, one of our runs (i.e. run 3) achieved the second best score reported in this year's KIS task.

6 Conclusions

In this paper we reported the ITI-CERTH framework for the TRECVID 2012 evaluation. ITI-CERTH participated in the SIN, KIS, MED and MER tasks in order to evaluate existing techniques and algorithms.

Regarding the TRECVID 2012 SIN task, many of the employed concepts are related to a significant motion pattern. In order to take advantage of the motion activity in each shot we have extracted 2-dimensional slices, named tomographs, with one dimension in space and one in time. The use of these

tomographs, as well as the provided ontology resulted to an improvement of 20% over the baseline approach.

As far as KIS task is concerned, the results reported were satisfactory and specific conclusions were drawn. First, the video-based representation assisted the users in the retrieval task and the ASR and metadata lemur modules were the most effective ones. On the other hand, visual concept retrieval or BoW didn't provide an added value.

Finally, as far as the TRECVID 2012 MED and MER tasks are concerned a number of efficient algorithms that exploit only limited audio and/or static visual information were evaluated providing satisfactory performance in terms of both detection accuracy and execution time.

7 Acknowledgements

This work was partially supported by the projects GLOCAL (FP7-248984), PESCaDO (FP7-248594) and LinkedTV (FP7-287911), which are funded by the European Commission.

References

[1] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proc. of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[2] 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.

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

[4] A. Moumtzidou, P. Sidiropoulos, S. Vrochidis, N. Gkalelis, and S. Nikolopoulos et al. ITI-CERTH participation to TRECVID 2011. In *Proc. TRECVID 2011 Workshop*. 9th TRECVID Workshop, Gaithersburg, MD, USA, December 2011.

[5] Alan F. Smeaton, Paul Over, and Wessel Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In Ajay Divakaran, editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin, 2009.

[6] J. Molina, V. Mezaris, P. Villegas, G. Tolias, and E. Spyrou et al. MESH participation to TRECVID2008 HLFE. 6th TRECVID Workshop, Gaithersburg, USA, November 2008.

[7] Steephane Ayache and Georges Queenot. Video Corpus Annotation using Active Learning. In *European Conference on Information Retrieval (ECIR)*, pages 187–198, Glasgow, Scotland, mar 2008.

[8] S. Possos and H. Kalva. Accuracy and stability improvement of tomography video signatures. In *Proc. of the 2010 IEEE International Conference on Multimedia and Expo*, pages 133–137. ICME 2010, Singapore, 19-23 July 2010.

[9] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of 4th Alvey Vision Conference*, pages 147–151, 1988.

[10] C. G. M. Snoek, K. E. A. van de Sande, X. Li, M. Mazloom, Y.-G. Jiang, D. C. Koelma, and A. W. M. Smeulders. The MediaMill TRECVID 2011 Semantic Video Search Engine. In *Proc. of TRECVID Workshop*, Gaithersburg, USA, December 2011.

[11] J. V. Gemert, C.J. Veenman, A. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006.

[13] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. High-level event detection in video exploiting discriminant concepts. In *Proc. 9th International Workshop on Content-Based Multimedia Indexing (CBMI 2011)*, pages 85–90, Madrid, Spain, June 2011.

[14] I. Tsampoulatidis, N. Gkalelis, A. Dimou, V. Mezaris, and I. Kompatsiaris. High-level event detection system based on discriminant visual concepts. In *Proc. of the 1st ACM International Conference on Multimedia Retrieval (ICMR)*, pages 68:1–68:2, New York, NY, USA, 2011. ACM.

[15] Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.

[16] S. Escalera, D. M.J. Tax, O. Pujol, P. Radeva, and R. P.W. Duin. Subclass problem-dependent design for error-correcting output codes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(6):1041–1054, June 2008.

[17] S. Escalera, O. Pujol, and P. Radeva. On the decoding process in ternary error-correcting output codes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(1):120–134, January 2010.

[18] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki. Linear subclass support vector machines. *IEEE Signal Processing Letters*, 19(9):575–578, September 2012.

[19] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. Mixture subclass discriminant analysis. *IEEE Signal Processing Letters*, 18(5):319–332, May 2011.

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

[21] R. Mertens, H. Lei, L. Gottlieb, G. Friedland, and A. Divakaran. Acoustic super models for large scale video event detection. In *Proc. of the 2011 joint ACM workshop on Modeling and representing events*, J-MRE '11, pages 19–24, New York, NY, USA, 2011. ACM.

[22] J.F. Bonastre, N. Scheffer, and D. et al. Matrouf. ALIZE/SpkDet: A state-of-the-art open source software for speaker recognition. 2008.

[23] 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.

[24] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki. Mixture subclass discriminant analysis link to restricted Gaussian model and other generalizations. *IEEE Trans. on Neural Networks and Learning Systems*, 24(1):8–21, January 2013.

[25] F. Song, D. Mei, and H. Li. Feature selection based on linear discriminant analysis. *Proc. 2010 Int. Conf. on Intelligent System Design and Engineering Application*, 1:746–749, October 2010.

[26] The lemur toolkit. `http://www.lemurproject.org/`.

[27] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1-2):89–108, 2002.

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

[29] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.

[30] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. John Wiley and Sons, 2nd edition, 1997.