ITI-CERTH participation to TRECVID 2014

Nikolaos Gkalelis[1], Foteini Markatopoulou[1,2], Anastasia Moumtzidou[1], Damianos Galanopoulos[1], Konstantinos Avgerinakis[1], Nikiforos Pittaras[1], Stefanos Vrochidis[1], Vasileios Mezaris[1], Ioannis Kompatsiaris[1], Ioannis Patras[2]

[1] Information Technologies Institute/Centre for Research and Technology Hellas, 6th Km. Charilaou - Thermi Road, 57001 Thermi-Thessaloniki, Greece
{gkalelis, markatopoulou, moumtzid, dgalanop, koafgeri, npittaras, stefanos, bmezaris, ikom}@iti.gr
[2] Queen Mary University of London, Mile end Campus, UK, E14NS
i.patras@qmul.ac.uk

Abstract

This paper provides an overview of the runs submitted to TRECVID 2014 by ITI-CERTH. ITI-CERTH participated in the Semantic Indexing (SIN), Event Detection in Internet Multimedia (MED), Multimedia Event Recounting (MER), and Instance Search (INS) tasks. In the SIN task, techniques are developed that combine floating-point local descriptors with binary local descriptors. In addition a multi-label learning algorithm is employed that captures the correlations among concepts. In the MED task, static and motion visual features as well as visual model vectors are extracted, and an efficient method combining a new kernel discriminant analysis (DA) technique and conventional LSVM is evaluated. In the MER subtask of MED the linear version of our DA method is combined with a model vector approach for selecting the key semantic entities depicted in the video and best describe the detected event. Finally, the INS 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 (also known as high-level feature extraction task - HLFE) under the MESH (TRECVID 2008) and K-SPACE (TRECVID 2007 and 2008) EU-funded research projects. In 2009 ITI-CERTH participated as a stand-alone organization in the SIN and Search tasks ([2]), in 2010 and 2011 in the KIS, INS, SIN and MED tasks ([3], [4]) and in 2012 and 2013 in the KIS, SIN, MED and MER tasks ([5], [6]) of TRECVID. 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: semantic indexing, event detection in internet multimedia, multimedia event recounting and instance search tasks. In the following sections we will present in detail the employed algorithms and the evaluation for the runs we performed in the aforementioned tasks.

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

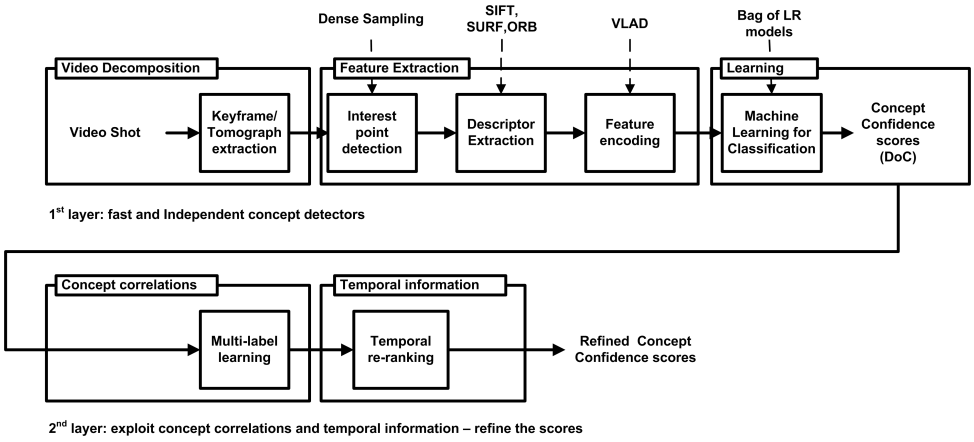

Figure 1: The general pipeline of our concept detection system.

2 Semantic Indexing

2.1 Objective of the submission

In TRECVID 2014, the ITI-CERTH participation in the SIN task [7] was based on an extension of our SIN 2013 system with new local descriptors and feature encoding approaches. The goal in this task is using the concept detectors to retrieve for each concept a ranked list of 2000 test shots that are mostly related with it. Our aim is to optimally combine the output of linear classifiers, based on multiple shot representations (keyframes, tomographs) and different local descriptors in order to improve system's performance, both in terms of accuracy and computational cost. To achieve our goal we enhance our SIN 2013 system with the following strategies:

- We introduce a binary local descriptor, namely ORB, which was originally proposed for similarity matching between local image patches, and we examine how effectively can be used in the task of semantic indexing.
- We introduce two color extensions of SURF. More specifically, based on a previously proposed paradigm for introducing color extensions of SIFT, we define in the same way color extensions for SURF, namely RGB-SURF and OpponentSURF.
- We use the VLAD encoding to aggregate different local descriptors of a video shot into a global image representation, in replacement of the Bag-of-Words approach.
- We employ a two-layer stacking architecture to capture concept correlations.

2.2 System Overview

Figure 1 shows the pipeline of the employed two-layer concept detection system. The first layer builds multiple independent concept detectors. The second layer takes as input the output of the first layer, exploits concept correlations and refines the initial scores.

Specifically, in the first layer of our system, the video stream is initially sampled, generating one keyframe and two visual tomographs per shot [8]. Subsequently, each sample is represented using one or more types of appropriate features (e.g., SIFT [9], SURF [10], ORB [11] etc.). These features are aggregated into global image descriptor vectors (using the VLAD encoding), which are given as input to a number of base classifiers (we use Logistic Regression (LR)) in order to build concept detectors that solve the problem of associating image descriptor vectors and concept labels. Then, when a new unlabeled video shot arrives, the trained concept detectors return confidence scores (a.k.a. Degree of Confidence - DoC) that show the belief of each detector that the corresponding concept appears in the shot. Finally, the outputs of all the concept detectors for the same concept are fused to estimate a final

concept detection score. It should be noted that this process is executed multiple times, independently for each one of the considered concepts that are to be detected.

In the second layer of the stacking architecture, the fused scores from the first layer are aggregated in model vectors and refined by two different approaches that work in cascade. The first approach uses a multi-label learning algorithm that incorporates concept correlations [12]. The second approach is a temporal re-ranking method that re-evaluates the detection scores based on adjacent video segments as proposed in [13].

The main features of our concept detection system are:

1. Seven local descriptors: SIFT and two color extensions of SIFT [14], namely RGB-SIFT and OpponentSIFT; SURF and two color extensions for SURF, inspired by the two color extensions of SIFT [14], namely RGB-SURF and OpponentSURF [15]; a binary local descriptor namely ORB (Oriented FAST and Rotated BRIEF) [11].

2. The aggregation of the local descriptors using the VLAD encoding.

3. A methodology for building concept detectors, where an ensemble of five LR models, called a Bag of Models (BoMs) in the sequel, is trained for each local descriptor and each concept [6].

4. The introduction of a multi-label classification algorithm in the second layer of the stacking architecture to capture label correlations [12].

2.2.1 Building Independent Concept Detectors

We developed image representations following the experimental setup of [16]. More specifically, we use the dense SIFT descriptor, that accelerates the original SIFT descriptor, in combination with the Pyramid Histogram Of visual Words (PHOW approach) [17]. PHOW is a simple modification of dense SIFT that uses more than one square regions at different scale levels in order to extract features. The same square regions at different scale levels of the PHOW approach are used as the image patches that were described by ORB and SURF. We calculate 128-SIFT, 128-SURF and 256-ORB grayscale descriptors; then, each color extension of SIFT and SURF descriptor results in a color descriptor vector three times larger than that of the corresponding original descriptor. All the local descriptors, except for the ORB, are compacted to 80 dimensions using PCA and are subsequently aggregated using the VLAD encoding [18]. Each image is divided into eight regions using spatial binning and sum pooling is used to combine the encodings from different regions. As a result of the above process, a VLAD vector of 163840 elements for SIFT or SURF and of 524288 elements for ORB is extracted for each image (by image we mean here either a keyframe or a visual tomograph). These VLAD vectors are compressed into 4000-element vectors by applying a modification of the random projection matrix [19]. These reduced VLAD vectors served as input to the Logistic Regression (LR) classifiers. Following the *cross validated committees* methodology of [6], we train five LR classifiers per concept and per local descriptor (SIFT, ORB, RGB-SIFT etc.), and combine their output by means of late fusion (averaging). When different descriptors are combined, again late fusion is performed by averaging of the classifier output scores.

2.2.2 Stacking for Exploiting Concept Correlations

We introduce a second layer in the concept detection pipeline in order to capture concept correlations. More specifically, we obtain concept score predictions from the individual concept detectors in the first layer, in order to create a *model vector* for each shot. These vectors form a meta-level training set, which is used to train a multi-label learning algorithm. We choose the Label Powerset (LP) [15] algorithm that models correlations among sets of more than two concepts. Our stacking architecture learns concept correlations in the second layer both from the outputs of first-layer concept detectors and by modelling correlations directly from the ground-truth annotation of a meta-level training set, which is a subset of the TRECVID 2014 annotated development set [20].

2.3 Description of runs

Four SIN runs were submitted in order to evaluate the potential of the aforementioned approaches on the TRECVID 2014 SIN dataset [21]. All 4 runs were based on building ensembles of LR models. We opted to investigate the following issues: a) whether a binary local descriptor (ORB) can complement the non-binary descriptors currently used in this task and b) if score refinement using the LP multi-label algorithm can increase the system's performance.

The 4 submitted runs for the main task of the 2014 TRECVID SIN competition are briefly described in the following:

- ITI-CERTH-Run1: "Multilabel2; Run using keyframes and tomographs as shot samples; SIFT and SURF color variants (SIFT, RGB-SIFT, OpponentSIFT, SURF, RGB-SURF, OpponentSURF) and binary ORB descriptors; VLAD encoding and dimensionality reduction via Random Projection variant; learning via linear logistic regression, *cross validated committees* method, stacking using Label Powerset, and temporal re-ranking". In this run six different non-binary local descriptors and one binary local descriptor are extracted for each keyframe or visual tomograph. We train five models using *cross validated committees* method for each of the seven local descriptors extracted from keyframes and the local descriptors extracted from the two visual tomographs per shot (we only extract SIFT, RGB-SIFT and OpponentSIFT from visual tomographs). Therefore 13 BoMs (or 65 LR models) per concept are generated. The scores returned from the 65 LR models are fused using the arithmetic mean. The fused scores per concept are refined in the second layer of the stacking architecture firstly, by applying the LP multi-label learning algorithm that captures concept correlations and secondly, by applying the temporal re-ranking method.

- ITI-CERTH-Run2: "Multilabel1; Run using keyframes and tomographs as shot samples; SIFT and SURF color variants (SIFT, RGB-SIFT, OpponentSIFT, SURF, RGB-SURF, OpponentSURF) descriptors; VLAD encoding and dimensionality reduction via Random Projection variant; learning via linear logistic regression, *cross validated committees* method, stacking using Label Powerset, and temporal re-ranking". This run is similar to the previous one ("Multilabel2"), the only difference being that here we use only non-binary descriptors, i.e., we do not include ORB in the set of descriptors that we use.

- ITI-CERTH-Run3: "Baseline; Run using keyframes and tomographs as shot samples; SIFT and SURF color variants (SIFT, RGB-SIFT, OpponentSIFT, SURF, RGB-SURF, OpponentSURF) descriptors; VLAD encoding and dimensionality reduction via Random Projection variant; learning via linear logistic regression, *cross validated committees* method, and temporal re-ranking". This is the baseline run that exacts only non-binary local descriptors. In this run we train five models using *cross validated committees* method for each of the six non-binary local descriptors extracted from keyframes and the local descriptors extracted from the two visual tomographs per shot (we only extract SIFT, RGB-SIFT and OpponentSIFT from visual tomographs). Therefore 12 BoMs (or 60 LR models) per concept are retrieved. Only temporal re-ranking is used for score refinement, i.e., the difference from the previous run "Multilabel1" is that here we do not use the LP step of the second layer.

- ITI-CERTH-Run4: "Binary; Run using keyframes as shot samples; binary ORB descriptor; VLAD encoding and dimensionality reduction via Random Projection variant; learning via linear logistic regression, *cross validated committees* method, and temporal re-ranking". This run uses only keyframes (no tomographs). We extract only ORB binary local descriptors from each keyframe (i.e. no other local descriptor was used in this run). Thus, one BoM (i.e. 5 LR classifiers) per concept are trained and are used. Temporal re-ranking is again used for score refinement.

2.4 Semantic Indexing Task Results

Table 1 summarizes the evaluation results of the aforementioned runs in terms of the Mean Extended Inferred Average Precision (MXinfAP). Moreover, in Table 2 we compare two runs in terms of the number of concepts for which a better performance was achieved. For instance in the first column

Table 1: Mean Extended Inferred Average Precision (MXinfAP) for all single concepts and runs.

	ITI-CERTH 1	ITI-CERTH 2	ITI-CERTH 3	ITI-CERTH 4
MXinfAP	0.207	0.205	0.190	0.081

Table 2: Number of improved concepts per run.

run4to3	run4to2	run4to1	un3to2	run3to1	run2to1
30	30	30	25	27	20

(run4to3: 30) we see that for 30 concepts Run 3 outperformed Run 4. From the obtained results the following conclusions can be drawn:

The "Binary" run (ITI-CERTH-Run-4), as expected, presents the lowest accuracy. The "Baseline" run (ITI-CERTH-Run-3) was used for assessing the use of non-binary local descriptors in combination with the VLAD encoding. The "Multilabel1" run (ITI-CERTH-Run-2) extends the "Baseline" run with a second layer that captures concept correlations and refines the concept detection scores. The higher performance achieved by this run demonstrates the significance of considering concept correlations in concept detection. This technique shows a relative improvement of 7.9% over the baseline run. Finally, the "Multilabel2" run (ITI-CERTH-Run-1) fuses binary and non-binary local descriptors and similarly to the "Multilabel1" run captures concept correlations. This technique shows that including ORB in the pool of descriptors does not improve significantly the overall concept detection accuracy; however, it does improve it slightly, and also slightly improves the XinfAP for 20 out of the 30 evaluated concepts.

It should be noted that after submission we discovered a bug in the normalization process of the VLAD vectors. Fixing this bug the MXinfAP for each run is expected to be increased by approximately 0.02.

Overall, considering our best run (0.207 MXinfAP), our system performs slightly above the median for 23 out of 30 evaluated concepts, as shown in Fig. 2. This good result is achieved despite the fact that we made design choices that favor speed of execution over accuracy (use of linear LR, dimensionality reduction of VLAD vectors).

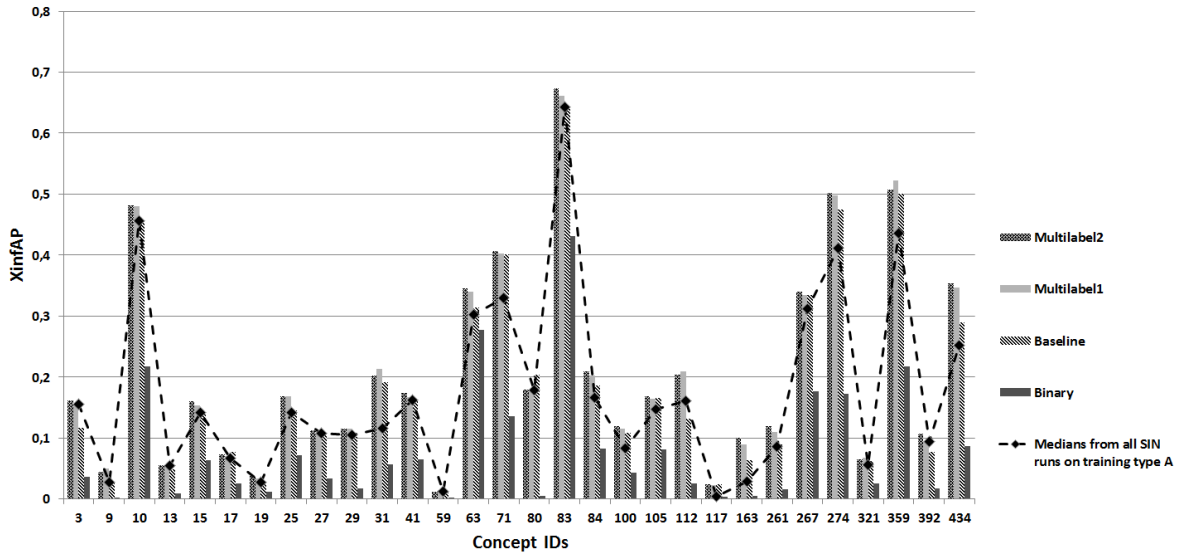

Figure 2: Extended Inferred Average Precision (XinfAP) per concept for our submitted runs.

3 Multimedia Event Detection and Recounting

3.1 Objective of the submission

One of the major problems in large-scale event detection (and in machine learning in general) is the high computational cost associated with learning the target events. This is the case in the MED/MER task, where usually a large number of high dimensional features are exploited in order to provide rich information for describing the different events depicted in video signals. For instance, in conventional PCs, both linear and kernel SVM (which are currently among the state-of-the-art classifiers) require very large training times to learn the MED target events and their descriptions for the MER task. This problem is more apparent in the Ad Hoc MED/MER task where a short period of time is only provided for extracting discriminative features and learning the target events. To this end, the objective of our submission is the evaluation of our detection system that combines a new fast discriminant analysis (DA) approach in combination with a conventional linear SVM (LSVM) [22] classifier.

3.2 System Overview

The target of the event detection and recounting 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 the event class (labelled with the integer one) or to the "rest of the world" class (labelled with the integer two). Additionally, for each positive detection, the system should produce a MER document recounting the key semantic entities $c_1, \ldots, c_I$ of the detected event depicted in the video. This is typically achieved using a training set $\mathcal{X} = \{\mathbf{X}_i^p | p = 1, \ldots, N_i,\ i = 1, 2\}$, where, $\mathbf{X}_i^p$ denotes the p-th video of the i-th class and N_i is the number of videos belonging to the i-th class, and a pool of concept detectors for capturing the semantic video content.

3.2.1 Metadata generation

Our method exploits three types of visual information, i.e., static, motion, and model vectors. For the extraction of static visual features and model vectors the procedure described in Section 2 is applied. We briefly describe the different visual modalities in the following:

- Each video is decoded into a set of keyframes at fixed temporal intervals (one keyframe every six seconds). Low-level feature extraction and encoding has been performed as described in Section 2. Specifically, four different local descriptors (SIFT, OpponentSIFT, RGB-SIFT, RGB-SURF), are applied to extract local visual information for every keyframe. The extracted features for each local descriptor are encoded using VLAD, compressed using a modification of the random projection matrix [19] technique to $\mathbb{R}^{4000}$, and averaged over all keyframes of the video. The four feature vectors are then concatenated to provide a single feature vector in $\mathbb{R}^{16000}$ at video level, encoding static visual information.

- A model vector representation of videos is created, similarly to [23, 24], in three steps: a) low-level feature extraction, b) evaluation of a set of external concept detectors at keyframe level, and, c) a pooling strategy to retrieve a single model vector at video level. Specifically, the four low-level feature representations at keyframe level described above are directly exploited. A pool of 1384 external concept detectors (346 concepts $\times$ 4 local descriptors), the ones derived in the SIN task, is then used to represent every keyframe with a set of model vectors (one model vector for each feature extraction procedure). The model vectors referring to the same keyframe are aggregated using the arithmetic mean operator, and subsequently, the model vectors of a video are averaged to represent the video in $\mathbb{R}^{346}$.

- For encoding motion information we use improved dense trajectories (DT) [25]. Specifically we employ the following four low-level feature descriptors: Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF) and Motion Boundary Histograms in both x (MBHx) and y (MBHy) directions. Hellinger kernel normalization is applied to the resulting feature vectors followed by Fisher Vector (FV) encoding with 256 GMM codewords. Subsequently, the four feature vectors are concatenated to yield the final motion feature descriptor for each video in $\mathbb{R}^{101376}$.

The final feature vector representing a video is formed by concatenating the feature vectors derived for each visual modality (static, motion, model vectors), yielding a new feature vector in $\mathbb{R}^{117722}$.

3.2.2 Event query generator

Event detection is accomplished using a nonlinear discriminant analysis (DA) method to derive a lower dimensional embedding of the original data, and a fast LSVM in the resulting subspace. The feature vectors derived with the procedure described in Section 3.2.1 by processing the MED 2014 positive and background videos, are used for training our event detectors.

In more detail, for dimensionality reduction we utilized a novel very fast DA method recently developed in our lab, based on our previous methods KMSDA and GSDA [26, 27, 28] called spectral regression kernel subclass discriminant analysis (SRKSDA), which is shown to outperform other DA approaches in both accuracy and computational efficiency (particularly at the learning stage). Given the positive and negative training data for a specific event, this DA method learns a transformation of the 117722-dimensional feature space in which the videos are originally represented, as described above, to a just D-dimensional space, $D \in [2, 3]$, which is discriminant for the specific event in question. Subsequently, a LSVM is trained as the event classifier in the resulting D-dimensional space. The SRKSDA has been implemented in Matlab [29], while for LSVM the libsvm [30] library is utilized.

This SRKSDA+LSVM learning process is in contrast to the usual approach of training a LSVM directly in the original, very high dimensional space. Experiments show that following our SRKSDA+ LSVM approach in most cases we achieve to learn more accurate event detectors in comparison to using a LSVM in the original, very high dimensional space, and at the same time the training of the SRKSDA+LSVM combination is completed in significantly less time than the time it would take to just train a LSVM in the original space, due to the dimensionality of that space (by significantly less time we mean here about 2 orders of magnitude shorter training time). Additionally, similar or better performance, and an even larger speed-up during the training stage, are observed in comparison to KSVM.

3.2.3 Multimedia event recounting

For each positively detected video an event recounting was additionally provided by our system. The event recounting was generated using the linear version of our new DA method (called SRSDA), the model vectors at keyframe level derived using our SIN 2014 concept detectors, and a semantic query for each event produced manually by visual inspection of the event description kit and our concept detectors.

SRSDA is employed as a feature selection method, similarly to the MSDA-based feature selection method, as described in [26, 27]. Using the above technique for each event a transformation matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2], \mathbf{w}_i \in \mathbb{R}^{346}$ is derived. The derived transformation matrix is then used for selecting the fifteen most discriminant concepts concerning the target event. This is done by firstly computing the weighted video model vector $\mathbf{y}^\tau = [y_1^\tau, \ldots, y_{346}^\tau]^T$ using

$$\mathbf{y}^\tau = \text{argmax}(\mathbf{w}_1 \circ \mathbf{x}^\tau, \mathbf{w}_2 \circ \mathbf{x}^\tau) \tag{1}$$

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 fifteen most discriminant concepts are then selected according the following rule

$$\{c_1, \ldots, c_{15}\} = \text{argmax}_I(y_1^\tau, \ldots, y_{346}^\tau)\,. \tag{2}$$

The MER output file is generated in the required format using the derived concepts $\{c_1, \ldots, c_{15}\}$, the associated concept detection scores, the event semantic query, and the DTD schema provided by NIST.

3.3 Dataset description

For training our PS and AH event detectors we used the PS-Training and AH-Training video sets consisting of 2000 (80 hours) and 1000 (40 hours) positive (or near-miss) videos respectively, and the

Table 3: Evaluation results (MAP) for the MED task.

	010Ex	100Ex
PS	15.1	30.3
AH	18.3	33.1

Event-BG video set containing 5000 (200 hours) of background videos. The PS and AH events were 20 and 10 respectively, and are listed below for the shake of completeness:

- PS events: "E021: Bike trick", "E022: Cleaning an appliance", "E023: Dog show", "E024: Giving directions", "E025: Marriage proposal", "E026: Renovating a home", "E027: Rock climbing", "E028: Town hall meeting", "E029: Winning race without a vehicle", "E030: Working on a metal crafts project", "E031: Beekeeping", "E032: Wedding shower", "E033: Non–motorized vehicle repair", "E034: Fixing a musical instrument", "E035: Horse riding competition", "E036: Felling a tree", "E037: Parking a vehicle", "E038: Playing fetch", "E039: Tailgating", "E040: Tuning a musical instrument".
- AH events: "E041: Baby Shower", "E042: Building a Fire", "E043: Busking", "E044: Decorating for a Celebration", "E045: Extinguishing a Fire", "E046: Making a Purchase", "E047: Modeling", "E048: Doing a Magic Trick", "E049: Putting on Additional Apparel", "E050: Teaching Dance Choreography".

For the evaluation of our system we processed the MED14–EvalSub set consisting of 32000 videos (960 hours). We submitted runs for both the 010Ex, 100Ex evaluation conditions (i.e., only 10 or 100 positive exemplars, respectively, are used for learning the specified event detector).

3.4 Description of runs

We submitted 4 runs in total. Specifically, we submitted our evaluation results for the PS and AH tasks, and within each task for the conditions 010Ex and 100Ex. That is the following runs were submitted: MED14Sub_PS_010Ex, MED14Sub_PS_100Ex, MED14Sub_AH_010Ex, MED14Sub_AH_100Ex.

In all runs the SRKSDA+LSVM method was used to build the event detectors and perform the event search in the MED14Sub set. During the learning stage a 3 cycle cross-validation procedure was employed. At each CV cycle the training set was divided to 70% learning and 30% validation set for learning the kernel parameters of SRKSDA. At the evaluation stage the detector of the specified event is applied to the evaluation set, providing a DoC for each video. The derived DoCs are ranked in descending order, and a detection threshold is computed for each event. Specifically, the threshold is selected as the $300th$ score of the ranked list for the 010Ex condition and as the $100th$ score for the 100Ex task. For the PS and AH tasks with 010Ex condition, for each video detected as a positive instance of the event, a MER document is additionally generated using the SRSDA-based feature selection method and the event semantic query.

3.5 Multimedia Event Detection and Recounting Results

The evaluation results of our 4 runs for the MED task are shown in Table 3, in terms of MAP along the 20 and 10 target events for the PS and AH tasks, and the 010Ex and 100Ex conditions. Moreover, our results for our MER evaluation along all events are depicted in Table 4. For this task the following measures defined in MER task were applied: i) "Query Conciseness", ii) "Key Evidence Convincing", and, iii) "Duration of Key Evidence Snippets" in percentage to the overall video duration. Taking into account the Likert texts provided by NIST, the actual evaluation for each MER document concerning the two first measures above was realized by five NIST judges and performed in terms of one of the following five predefined measurement scales: a) Strongly Disagree, b) Disagree, c) Neutral, d) Agree, e) Strongly Agree.

From the analysis of the evaluation results we can conclude the following:

Table 4: Evaluation results for the MER task.

	Query Conciseness	Key Evidence Convincing
Agree or Strongly Agree	18%	52%
Disagree or Strongly Disagree	72%	34%
Neutral	10%	13%

Duration of Key Evidence Snippets: 9.1%

- In comparison to most of the other submissions we still employ only a small number of visual features. Nevertheless, among the submissions that processed only the MED14–EvalSub set our system provides the best performance. Moreover, considering all submissions to the AH task, our submission is among the seven best (in both the 010Ex and 100Ex conditions).
- Concerning our previous year submission we observed a large performance gain (in terms of MAP) of more than 12% and 20% in the AH and PS tasks respectively. This is due to improvements at all levels of our system, such as, the use of improved visual static feature descriptors and concept detectors, exploitation of motion information, application of the new DA preprocessing step to extract the most discriminant features, and an overall application of a faster detection method that allows for a more effective optimization.
- Our approach is particularly suited for Ad Hoc event detection task for the following reasons: a) relies on learning automatically from only the video training data which dimensions of the original very high-dimensional feature space can contribute to the building of a very low-dimensional discriminant feature space, without the need to e.g. manually select which concepts are more or less important for a given event, b) using the fast DA approach, the training procedure of several event detectors and recounters can be completed in a short period of time, as required in the AH task.
- Concerning the MER task, from the attained results we conclude that the overall performance of our run is average to low. This was rather expected because only a small set of semantic concepts are used, which additionally exploit only limited static visual information (motion information is not currently exploited by our concept detectors). However, for the "Key Evidence Convincing" measure a rather good performance is attained, where in 52% of the cases the judges "agree" or "strongly agree" with the key evidences described in the submitted MERs, and in 13% have a neutral opinion. Moreover, we should note that a very low "Key Evidence Snippets Duration" is observed, which facilitates examination and further exploitation of the MERs. This is due to the fact that we localize the key evidences at keyframe level (in contrary to our previous year submission where we provided evidences at video level).

4 Instance Search

4.1 Objective of the submission

ITI-CERTH's participation in the TRECVID 2014 Instance Search (INS) task aimed at studying and drawing conclusions regarding the effectiveness of different retrieval modules, which are integrated in VERGE[1] interactive video search engine, in the retrieval procedure. According to the TRECVID guidelines, the INS task represents the situation, in which the user is searching for video segments of a specific person, object, or place contained in a video collection. It should be noted, that the searcher is provided with visual examples of the specific query object in order to commence with the searching [21].

Two runs are submitted that differ exclusively in the temporal segments used for video representation, i.e. shots and scenes. Finally, it should be noted that the videos used in the INS task are provided by BBC and they are part of the EastEnders TV series (Programme material BBC).

4.2 System Overview

The system employed for the Instance Search task was VERGE, 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. 3. The following modules are integrated in the developed search application:

- Visual Similarity Search Module;
- Transcription and Metadata Search Module;
- High Level Visual Concept Retrieval;

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

VERGE can support two modes of video representation a) shot-based representation and, b) scene-based representation. The shots of the video are provided by the organizers, while scenes are recognized using the algorithm introduced in [31] that groups the shots into scenes based on the visual similarity and the temporal consistency among them. In general, the scene segmentation module is applied in order to allow the system to deal with the large amount of data provided, to aid the data indexing and improve the scalability of the search engine.

It should be noted, that the search system is built on open source web technologies and more specifically the HTTP Apache server, Apache Solr, PHP, JavaScript and mySQL database.

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

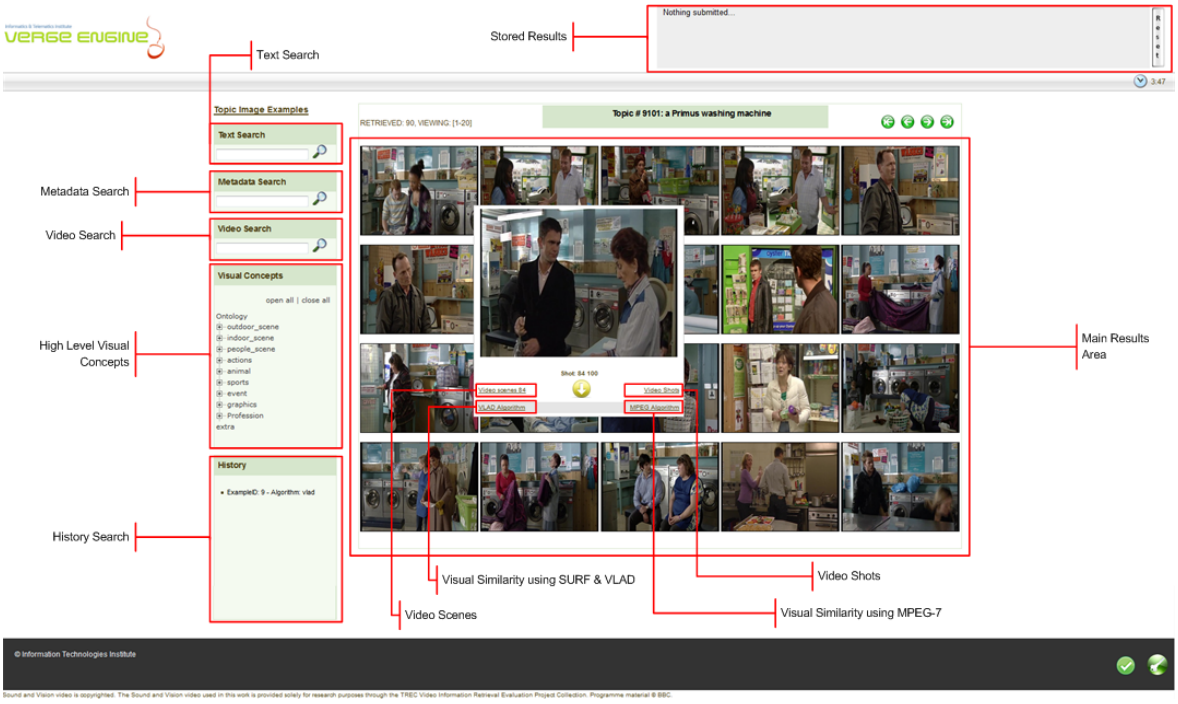

Figure 3: User interface of the interactive search platform.

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

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.

Visual similarity is realized using global and local information. To deal with global information, MPEG-7 descriptors (i.e. Colour Layout and Edge Histogram) are extracted from each keyframe and they are concatenated into a single feature vector. Regarding the case of local information, SURF features are extracted, then K-Means clustering is applied on the database vectors in order to acquire the visual vocabulary and finally VLAD encoding is realized for representing images [18].

For the Nearest Neighbour search, we create an Asymmetric Distance Computation index and then, we compute the K-Nearest Neighbours from the query image. It should be noted that this indexing structure is utilized for both descriptors (i.e. global and local). Finally, web services are implemented for both descriptors and are used for accelerating the querying and retrieval process. This is achieved since in order to query the indexing structures, a two-step procedure is realized that involves: a) the loading of the index on the RAM memory, and b) the querying to the index. The first one of these steps is very time-consuming since it requires more than three minutes. Therefore, in order to eliminate the time required for the index loading, web services are created that load these indexing structures on the RAM memory at the beginning of all runs, and thus allows instant querying of the structures and thus fast results retrieval each time a query is realized.

4.2.2 Transcription and Metadata Search Module

The transcription search 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 [20]. Regarding the metadata module, it exploits the information (i.e. cast and synopsis data) provided along with every video of the collection.

In both modules, a pre-processing step is initially applied that involves the processing of the acquired content and includes punctuation and stop words removal. Finally, the processed content is indexed using Apache Solr search platform[2] that allows full-text search and enables fast retrieval as well easy formulation of complicated queries.

Finally, it should be noted that in case of the transcription module and when the scene representation mode is used, the scene associated text consists of the concatenation of the transcriptions of all included shots.

4.2.3 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. The concepts that are incorporated into the system are the 346 concepts studied in the TRECVID 2014 SIN task using the techniques and the algorithms described in detail in section 2.

4.3 Instance Search Task Results

The system developed for the Instance Search task includes the aforementioned modules. We submitted two runs to the INS task that differ on the video representation mode used (i.e. the first assumes shot-based representation, while the second scene-based representation), while the search modules described earlier are available in both runs. It should be mentioned that the complementary functionalities (i.e. shot-segmented view of each video, storage structure and history bin) were available in both runs. According to the TRECVID guidelines the time duration for each run was set to fifteen minutes. Regarding the users that participated in INS runs, a latin square design was applied and thus six users were engaged in total and each one run 8 topics, therefore devoted 2 hours to the system evaluation. In general, although the educational and age profile of the users is similar, their experience with similar systems differs (e.g some users were expert users while others were moderate users). The mean average precision as well the recall for all runs are illustrated in Table 5.

[2]Apache Solr: http://lucene.apache.org/solr/

Table 5: Evaluation of search task results.

Run IDs	Mean Average Precision	Recall
I_NO_ITI_CERTH_1 (Shot_based representation)	0.032	532/9336
I_NO_ITI_CERTH_2 (Scene_based representation)	0.028	315/9336

By comparing the values of table 5, we can draw conclusions regarding the effectiveness of the representation modes given that the search techniques used, are the same for both runs. Specifically, it becomes clear that the shot-based representation mode outperforms the scene-based. Therefore given that the response time is the same for both modes and the computer can handle the memory required for the shot-based mode, there is no apparent reason for preferring the scene mode that decreases significantly the information captured by the system. Finally, another conclusion that was drawn using the user opinions and feedback after their interaction with the system, but cannot emerge from table 5, is that the textual modules (i.e. both the ASR data and the metadata) weren't of any use for them.

Finally, it should be noted that the system developed, although achieved a rather low performance compared to the other systems competing in the INS task, it has shown significant improvement compared to the last year results.

5 Conclusions

In this paper we reported the ITI-CERTH framework for the TRECVID 2014 evaluation [21]. ITI-CERTH participated in the SIN, INS, MED and MER tasks in order to evaluate new techniques and algorithms.

Regarding the SIN task, various techniques were developed and combined reporting good results. VLAD encoding was used to aggregate different local descriptors; new color extensions of SURF were combined with state-of-the-art floating-point local descriptors and one binary local descriptor, namely ORB; a method that exploits concept correlations was used as part of a stacking-based architecture.

Concerning the MED and MER tasks a new algorithm, combining discriminant analysis and linear SVM, was evaluated providing good performance in terms of both accuracy and response time.

Finally, as far as INS task is concerned, the results reported were far better that last year results but there is still a lot of room for improvement in order for the system to become competitive against the other systems. The most important conclusions from this year runs was that the scene-based representation mode is inferior to the shot-based one and thus it will not be used in following systems, and that the textual information is not very useful in the context of INS task.

6 Acknowledgements

This work was partially supported by the European Commission under contracts FP7-287911 LinkedTV, FP7-600826 ForgetIT, FP7-610411 MULTISENSOR and FP7-312388 HOMER.

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] A. Moumtzidou, N. Gkalelis, P. Sidiropoulos, M. Dimopoulos, and S. et al. Nikolopoulos. ITI-CERTH participation to TRECVID 2012. In *TRECVID 2012 Workshop*, Gaithersburg, MD, USA, 2012.

[6] F. Markatopoulou, A. Moumtzidou, C. Tzelepis, K. Avgerinakis, N. Gkalelis, S. Vrochidis, V. Mezaris, and I. Kompatsiaris. ITI-CERTH participation to TRECVID 2013. In *TRECVID 2013 Workshop*, Gaithersburg, MD, USA, 2013.

[7] A. F. Smeaton, P. Over, and W. 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.

[8] P. Sidiropoulos, V. Mezaris, and I. Kompatsiaris. Video tomographs and a base detector selection strategy for improving large-scale video concept detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(7):1251–1264, July 2014.

[9] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.

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

[11] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *IEEE Int. Conf. on Computer Vision*, pages 2564–2571, 2011.

[12] F. Markatopoulou, V. Mezaris, and I. Kompatsiaris. A comparative study on the use of multi-label classification techniques for concept-based video indexing and annotation. In Cathal Gurrin, Frank Hopfgartner, Wolfgang Hurst, Hvard Johansen, Hyowon Lee, and Noel OConnor, editors, *MultiMedia Modeling*, volume 8325 of *Lecture Notes in Computer Science*, pages 1–12. Springer International Publishing, 2014.

[13] B. Safadi and G. Quénot. Re-ranking by Local Re-Scoring for Video Indexing and Retrieval. In C. Macdonald, I. Ounis, and I. Ruthven, editors, *CIKM*, pages 2081–2084. ACM, 2011.

[14] K. E. A. Van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1582–1596, 2010.

[15] F. Markatopoulou, N. Pittaras, O. Papadopoulou, V. Mezaris, and Patras I. A study on the use of a binary local descriptor and color extensions of local descriptors for video concept detection. In *Proc. 21th Int. Conf. on MultiMedia Modeling (MMM'15)*. Springer International Publishing, 2015.

[16] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, pages 76.1–76.12. British Machine Vision Association, 2011.

[17] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *IEEE Int. Conf. ICCV 2007*, pages 1–8, Rio de Janeiro, 2007.

[18] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012.

[19] E. Bingham and H. Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 245–250, NY, 2001. ACM.

[20] S. Ayache and G. Quenot. Video Corpus Annotation using Active Learning. In *European Conference on Information Retrieval (ECIR)*, pages 187–198, Glasgow, Scotland, 2008.

[21] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot. Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2014*. NIST, USA, 2014.

[22] V. Vapnik. *Statistical learning theory*. New York: Willey, 1998.

[23] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE Trans. Multimedia*, 14(1):88–101, February 2012.

[24] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. High-level event detection in video exploiting discriminant concepts. In *Proc. CBMI*, pages 85–90, Madrid, Spain, June 2011.

[25] H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013.

[26] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki. Video event recounting using mixture subclass discriminant analysis. In *IEEE International Conference on Image Processing, ICIP 2013, Melbourne, Australia, September 15-18, 2013*, pages 4372–4376, 2013.

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

[28] N. Gkalelis and V. Mezaris. Video event detection using generalized subclass discriminant analysis and linear support vector machines. In *International Conference on Multimedia Retrieval, ICMR '14, Glasgow, United Kingdom - April 01 - 04, 2014*, page 25, 2014.

[29] *MATLAB, User' s Guide*. The MathWorks, Inc., 1994-2001, `http://www.mathworks.com/`.

[30] C.C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2, 2011.

[31] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177, August 2011.