PicSOM Experiments in TRECVID 2006

Mats Sjöberg, Hannes Muurinen, Jorma Laaksonen
Adaptive Informatics Research Centre
Helsinki University of Technology, Finland

Markus Koskela
Centre for Digital Video Processing
Dublin City University, Ireland

Abstract

Our experiments in TRECVID 2006 include participation in the shot boundary detection, high-level feature extraction, and search tasks, using a common system framework based on multiple parallel Self-Organizing Maps (SOMs). In the shot boundary detection task we projected feature vectors calculated from successive frames on parallel SOMs and monitored the trajectories to detect the shot boundaries. We submitted the following ten runs:

- `PicSOM_CA`: cut-optimized using all the training videos
- `PicSOM_GA`: gradual-optimized using all the training videos
- `PicSOM_BA`: optimized for both cuts and gradual transitions using all the training videos
- `PicSOM_CN`: cut-optimized using only the news videos (without the NASA videos)
- `PicSOM_GN`: gradual-optimized using only the news videos
- `PicSOM_CS`: cut-optimized using channel-specific training videos
- `PicSOM_GS`: gradual-optimized using channel-specific training videos
- `PicSOM_CNF`: cut-optimized using only the news videos and only a few features
- `PicSOM_CNE`: cut-optimized using only the news videos and one additional edge feature
- `PicSOM_CAE`: cut-optimized using all the training videos and one additional edge feature

The trajectory-based method seemed to work comparatively well in the task. By comparing the F1 scores of the runs we found out that the results mostly degraded when using only a portion of the data in training. Especially the channel-specific detectors seemed to suffer from overfitting and did not work well probably because of the low amount of channel-specific training data compared to the number of adjustable parameters.

In the high-level feature extraction task, we applied a method of representing semantic concepts as class models on a set of parallel SOMs, combined with an inverse file created from automated speech recognition and machine translation (ASR/MT) data. We submitted six runs as follows:

- `A_SOM_F3_6`: still-image and video features
- `A_SOM_F4_5`: visual features and ASR/MT data
- `A_SOM_F5_4`: visual features and stemmed ASR/MT data
- `A_SOM_F6_3`: visual features and ASR/MT and closed-caption data
- `B_PicSOM_F7_2`: visual features and LSCOM concepts
- `B_PicSOM_F9_1`: visual features, ASR/MT data and LSCOM concepts

We observed increase in performance when adding both textual features and the auxiliary concepts to the visual features baseline.

In the search task, we submitted a total of six runs (five automatic and one interactive run). Our method used SOM and inverse file indices from visual and textual features combined with class models of appropriate semantic concepts. The overall settings for the runs were as follows:

- `F_A_1_OM-f1_6`: baseline automatic run using only ASR/MT data
- `F_A_2_OM-f2_5`: automatic run using only visual features
- `F_A_2_OM-f3_4`: automatic run using ASR/MT data and visual features
- `F_B_2_OM-f4_2`: automatic run using ASR/MT data, visual features, and LSCOM concepts
- `F_B_2_OM-f5_3`: automatic run using either only ASR/MT data or visual features and LSCOM concepts, selected by named entity detection
- `I_B_2_OM-i_1`: interactive run with ASR/MT data, visual features, and LSCOM concepts

Using class models created from the LSCOM concepts improved the retrieval performance as measured by MAP scores. Also the entity detection in the last automatic run proved successful and seems to be a promising topic for future experiments.

I. Introduction

In this paper, we describe our experiments with the PicSOM system in TRECVID 2006. We participated in the shot boundary detection, high-level feature extraction, and automatic and interactive search tasks.

In 2005, the first year that we participated in TRECVID [1], we showed that our existing system for indexing multimodal hierarchical objects using Self-Organizing Maps (SOMs) and relevance propagation was suitable for digital video retrieval, and the results compared promisingly with other systems. The basic system and methodology used this year is partially the same as in TRECVID 2005, but we also tested some evolutionary improvements to our old methods in combination with some new ideas and concepts. Many of these proved to have a positive impact on performance. When using only visual features and ASR/MT data, the 2006 system is comparable to the one used in 2005, although this year we had an extended set of visual descriptors. By comparing the results of these two years, we judge that the search tasks were more difficult this year than in the previous year.

A completely new task for us this year was shot boundary detection, where we proposed a novel idea based on tracking trajectories of projected feature vectors on parallel SOMs. The results were quite promising.

In the high-level feature extraction task, we applied our method of representing semantic concepts as *class models* on a set of parallel feature indices. This year we also utilized positive and negative *auxiliary* concepts from the Large Scale Concept Ontology for Multimedia (LSCOM) ontology [2]. In the search task, we used the LSCOM concepts as well, by matching of synonymous words using WordNet [3], in addition to textual and visual features.

The rest of the paper is organized as follows. The PicSOM system for video retrieval and the used visual and textual content descriptors are briefly described in Section II. A novel solution to the shot boundary detection task is described in Section III. Our experiments for the high-level feature extraction and search tasks are described in Sections IV and V, respectively, and conclusions are presented in Section VI.

II. Indexing Video with PicSOM

The PicSOM system [4] is a general framework for research on content-based indexing and retrieval of visual objects. The system is based on using several complementary Self-Organizing Maps (SOMs) [5], each trained with separate feature data. The SOM defines an elastic, topology-preserving grid of points that is fitted to the input space. The distribution of the data vectors over the map forms a two-dimensional discrete probability density. As a result, the different SOMs impose different similarity relations on the objects. The task of the retrieval system then becomes to select, weight and combine these similarity relations so that their composite would approximate the human notion of similarity in the current retrieval task as closely as possible. The parallel SOMs can also be augmented with other types of additional information and different indices. For the TRECVID 2006 evaluations, such a source of information is the ASR/MT text output, for which the inverted file provides an effective indexing structure.

Ordinary retrieval usage of the PicSOM system is based on relevance feedback: the user determines the relevance of all returned objects and marks the ones she considers relevant to the current task; the others are deemed non-relevant. The SOM units on all maps are awarded positive and negative scores for every relevant and non-relevant object mapped in them, respectively. The system remembers all responses the user has given since the query was started in these sparse value fields.

Due to the topology preservation property of the SOM, we are also motivated to spread this relevance information to the neighboring map units on the SOM grids. Spreading of the response values can be performed by convolving the sparse value fields with a tapered kernel function. This results in polarization of the entire map surface in areas of positive and negative cumulative relevance.

By locating a given database object in all SOM indices, we get its relevance scores with respect to the different features. Then, as the response values of the parallel indices are mutually comparable, we can determine a global ordering and the overall best candidate objects using simple unweighted linear combination.

To support multimodal fusion, e.g. between the different modalities of the video objects (visual, aural, textual), a hierarchical approach has been employed [6]. The multimodal hierarchy for video shots used for indexing the TRECVID 2006 collection is illustrated in Fig. 1. The video shot itself is considered as the main or parent object in the tree structure. The key frames (one or more) associated with the shot, the audio track, and ASR/MT text are linked as children of the parent object. All object modalities may have one or more SOMs or other feature indices, and thus all objects in the hierarchy may have links to a set of associated feature indices.

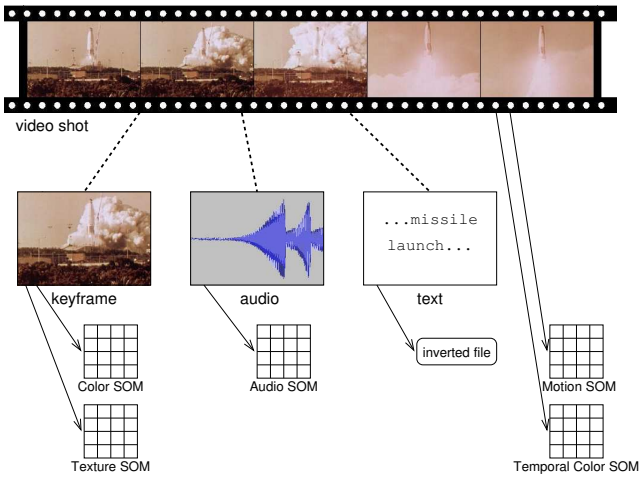


Fig. 1. The hierarchy of video and multimodal SOMs.

In indexing the video shots of the TRECVID 2006 collection, we used in total six video features and ten still image features. An aural feature was also tried, but later discarded as it did not perform particularly well. A separate 256×256-sized SOM was trained for each of these features. For the ASR/MT output, we used a concept-wise text feature based on an inverted file in the high-level feature extraction task. For the search task an inverted file of all the words used were created. All these features are briefly described below.

A. Video features

On the video shot level, we used the MPEG-7 [7] *Motion Activity* descriptor (MA) and temporal versions of five still image features. The temporal features were based on the following still image features (which are not MPEG-7): *Average Color*, *Color Moments*, *Texture Neighborhood*, *Edge Histogram* and *Edge Co-occurrence*. The still-image features are calculated for five spatial zones for each frame of the video clip. These results are averaged over the frames contained within each one of five non-overlapping temporal video slices. In this way we get a final feature vector that describes the changes of the still image descriptors over time in different spatial areas of the video.

The Average Color feature vector is a three-element vector that contains the average RGB values of all the pixels within the zone. The Color Moments feature treats the HSV color channels from the zone as probability distributions, and calculates the first three moments (mean, variance and skewness) are calculated for each distribution.

The Texture Neighborhood feature is calculated from the Y (luminance) component of the YIQ color representation of the zone pixels. The 8-neighborhood of each inner pixel is examined, and a probability estimate is calculated for the probabilities that the neighbor pixel in each surrounding relative position is brighter than the central pixel. The feature vector contains these eight probability estimates. Edge Histogram, is the histogram of four Sobel edge directions. It is not the same as the MPEG-7 descriptor with the same name. Edge Co-occurrence gives the co-occurrence matrix of four Sobel edge directions.

B. Image features

For the key frame indices we used a set of five standard MPEG-7 descriptors; *Color Layout*, *Color Structure*, *Dominant Color*, *Scalable Color* and *Edge Histogram*. The descriptors were extracted globally from every key frame in the collection, i.e. no segmentation or zoning was used.

In addition to the MPEG-7 features, we also used the same non-standardized still image features that were used with the temporal video descriptors. These were calculated for the five spatial zones of each image and the values concatenated to one image-wise vector.

C. Text features

Unlike the other features, an inverted file instead of a SOM index was used for the ASR/MT output. The extension of the PicSOM system for using such indices in parallel with the SOMs was presented in [8].

For the high-level feature extraction task, the text features were constructed by gathering concept-dependent lists of most informative terms. Let us denote the number of shots in the development set associated with concept c as N_c and assume that of these shots, $n_{c,t}$ contain the term t in the ASR/MT output. After preprocessing and stemming, the following measure is applied for term t regarding the concept c:

$$S_c(t) = \frac{n_{c,t}}{N_c} - \frac{n_{all,t}}{N_{all}} \ ,$$

where N_{all} is the total number of shots in the whole development set, and $n_{all,t}$ is the number of those that contain the term t. For every concept, we record the 100 most informative terms and use them as alternative text features.

III. Shot Boundary Detection

We participated in the shot boundary detection task for the first time this year. Our approach in this task was to use the topology preservation properties of SOMs

in spotting the abrupt and gradual transitions. Multiple feature vectors calculated from consecutive frames were projected on two-dimensional feature-specific SOMs. The transitions are detected by observing the trajectories formed on the maps.

Due to the topology preservation, similar inputs are mapped close to one another on the SOMs. The trajectory of the best-matching map units of successive frames thus typically hovers around some region during a shot, if the visual content of the video measured by the feature vectors does not change too rapidly. Abrupt cuts are characterized by sudden trajectory leaps from one region on the map to another, and gradual transitions on the other hand are characterized by a somewhat rapid drift of the trajectory from one region to another. Our detector tries to detect these kind of characteristic phenomena.

To increase detector robustness and prevent false positive cut detection decisions e.g. due to flashlights, we do not only monitor the rate of change of the map position between two consecutive frames, but take small frame windows from both sides of the current point of interest, and compare the two frame windows. A circular area with a constant radius is placed over each map point in the given frame window as illustrated in Figure 2. We call the union of these circular areas the area spanned by the frame window. If the areas spanned by the preceding and following frame windows overlap, there are some similar frames in both of them, and we decide that the current point of interest is not a boundary point. If there is no overlapping, the frames in the frame windows are clearly dissimilar, and we decide that we have found a boundary. The flashlights are characterized by sudden trajectory leaps to some region on the map followed by a leap back to the original region. If the duration of the flashlight is smaller than the frame window size, the proposed method helps to avoid false positives.

The final boundary decision is done by a committee machine that consists of this kind of parallel classifiers. There is one classifier for each feature calculated from the frames, and each classifier has a weight value. The final decision is made by comparing the weighted vote result of the classifiers to a threshold value. Abrupt cuts and gradual transitions are detected using the same method. The detected boundary points that are close to one another are combined, and as the result we get the starting locations and lengths of the transitions. To facilitate detection of slow gradual transitions, our system also allows to use a frame gap of given length between the two frame windows. A more detailed description of the algorithm is given in [9].

TABLE I

AN OVERVIEW OF THE SHOT BOUNDARY DETECTION TASK RUNS.

#	Run id	Videos All	Videos News	Videos Ch.	Features -	Features Norm.	Features +	Trans. Cut	Trans. Grad.
1	PicSOM_CA	•				•		•	
2	PicSOM_GA	•				•			•
3	PicSOM_BA	•				•		•	•
4	PicSOM_CN		•			•		•	
5	PicSOM_GN		•			•			•
6	PicSOM_CS		•	•		•		•	
7	PicSOM_GS		•	•		•			•
8	PicSOM_CNF		•		•			•	
9	PicSOM_CNE		•				•	•	
10	PicSOM_CAE	•					•	•	

A. *Parameter and feature selection*

The parameters of the detector like the circle radii, the window lengths, the vote threshold value and the frame gap lengths were selected using a discrete gradient descent method. We submitted ten runs, each of them using slightly different sets of training data in the training phase. Table I summarizes the training data used in the runs. The table specifies what portion of the training videos, which features, and which transitions of the videos were used to tune the parameters. In some runs we used all the video data, in some others we omitted the NASA videos and used only the news videos. We also tested channel-specific parameter tuning. In those runs we trained a separate detector for each channel in the training data, and used them with corresponding channels in the test data. A detector trained with all the news data was used for those channels in the test data that were not represented in the training data.

The feature column of Table I specifies whether a normal, a reduced, or an extended feature set was used. The normal set includes the ten abovementioned still image features calculated without segmentation or zoning. Automatic feature selection and weighting was done among these by tuning the feature weight parameters. The reduced feature set contains five features that gained the highest weights in the corresponding training phase with all the features. The extended feature set contains one additional edge feature, *Edge Fourier*, which was implemented after the other runs were already done. The additional feature vector contains the values of the 16x16 Fast Fourier Transformation of the Sobel edge image of the given frame. The transition column specifies what portion of the ground truth data was used in the training: only the cuts, only the gradual transitions, or both.

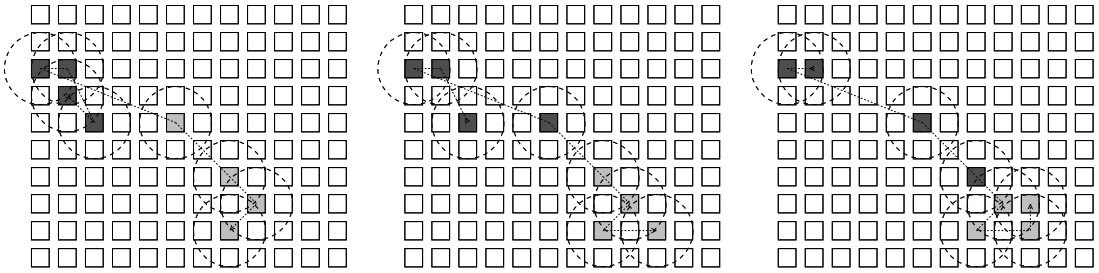


Fig. 2. Segments of a trajectory at three consecutive time steps. The SOM cells marked with dark gray color represent trajectory points belonging to the set of preceding frames, and light gray cells represent the following frames. The trajectory is illustrated with a dotted black line, and the circles represent the area spanned by the preceding and following frame sets. The areas do not overlap in the two leftmost figures, but overlap at the third figure. This is interpreted as a one frame gradual transition.

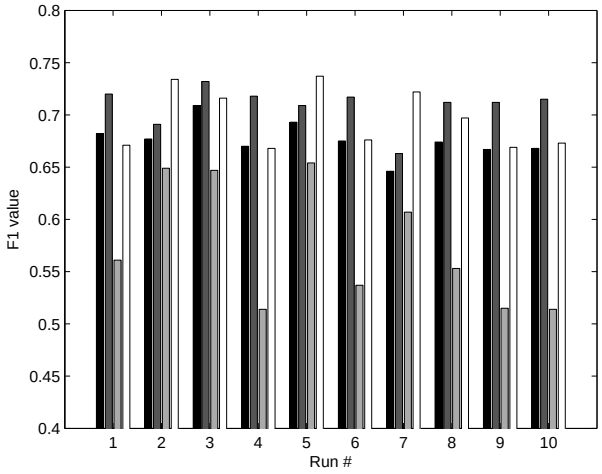


Fig. 3. The F1 values depicting the shot boundary detector performance: total performance (black), cut detection (dark gray), gradual detection (light gray), and F1 calculated from frame precision and frame recall (white).

B. Results

Figure 3 shows the F1 values calculated from the shot boundary detection results for each run. The gradual-optimized detectors 2, 5 and 7 seem to detect gradual transitions better than the cut-optimized detectors, but they seem to detect the cuts fairly well too. The gradual-optimized detectors typically have higher recall and lower precision values than the cut-optimized, since the detectors need to be more sensitive in order to detect the gradual transitions.

Omitting the NASA videos somewhat deteriorates the cut-optimized performance, but slightly improves the F1 values of the gradual-optimized detector, as can be seen by comparing detectors 1 and 2 to detectors 4 and 5 respectively. There are lots of slow gradual transitions in the NASA videos which are usually not found in news videos, and this might make the detector too sensitive. The reduction of cut-detection performance can probably be explained as overlearning. The low performance values of the channel-specific gradual-optimized classifier 7 is probably also caused by overlearning because there was not much training data for the channels. The cut-optimized channel-specific detector 6 detects gradual transitions slightly better than detector 4 trained with news data, but this is probably just a coincidence, since the gradual transitions were not used in training. Detector 1 trained with all the videos outperforms both of them.

Run number 3 seems to have the best F1 score for total and cut detection performance. The gradual transition detection performance is also comparable with the efficiency of the gradual-optimized detectors. This is the only run in which both the gradual and cut ground truth transitions were used in training, and thus this run had the largest total amount of training data. Surprisingly the cut detection performance improves when the parameter optimizer has to make compromises between cut and slow transition detection. We can also see from the results of runs 8, 9 and 10 that reducing the number of features slightly improves the overall performance, while including the extra feature decreases the performance. The number of parameters is increased as the number of features increase. These results support further the hypothesis of overlearning when having too many adjustable parameters compared to the amount of training data.

IV. High-level Feature Extraction

For the high-level feature extraction task, we used a similar approach as in our TRECVID 2005 submission [1]. The method is based on modeling probability densities of the concepts using kernel-based estimation of discrete class densities over the SOM grids. All 39 LSCOM-Lite [10] concepts are detected using the same

TABLE II
AN OVERVIEW OF THE RUNS IN THE HIGH-LEVEL FEATURE EXTRACTION TASK.

#	run id	image	video	nonst.	stem	cc	aux.
F1	not submitted	•					
F2	not submitted		•				
F3	A_SOM_F3_6	•	•				
F4	A_SOM_F4_5	•	•	•			
F5	A_SOM_F5_4	•	•		•		
F6	A_SOM_F6_3	•	•	•		•	
F7	B_PicSOM_F7_2	•	•				•
F8	not submitted	•	•	•		•	•
F9	B_PicSOM_F9_1	•	•	•			•

procedure based on the concept-wise ground-truth annotations. The method has proven to be readily scalable to large-scale concept ontologies, and this enabled us to model the concepts in the full LSCOM [2], [11] ontology in a similar manner and perform experiments in which we added auxiliary LSCOM concepts to augment the submitted 39 concept detectors.

We designed a set of nine high-level feature extraction experiments, and submitted a total of allowed six runs from them. The non-submitted runs (F1, F2, and F8) do not have run identifiers given by NIST so in the following discussion we refer to the performed runs using their corresponding run numbers (F1–F9). Table II gives an overview of the runs. The columns in Table II, listing the different information sources, refer to image and video features, non-stemmed and stemmed ASR/MT output, closed caption data, and auxiliary LSCOM concepts. Our sole submission to the high-level feature extraction task in TRECVID 2005 corresponds mostly to run F5, although we now had more visual features available.

In the first three runs, we establish a baseline of using only visual descriptors with only still-image features (run F1), only video features (run F2) and both of these modalities (run F3). For each concept, the subset of features actually used are selected as described in Section IV-A below. The next three runs explore the effect of different textual features along with the visual features. In these runs, the ASR/MT text features are included either as non-stemmed (run F4) or stemmed (run F5). In run F6, we combine the closed caption data with the non-stemmed ASR/MT text features.

In the last three runs, we include a selected set of auxiliary concepts for each concept detector. This technique is described in Section IV-B in more detail. In run F7, we compare the inclusion of the auxiliary concepts to the visual features baseline (run F3). Runs F8 and F9 incorporate also the text features (non-stemmed ASR/MT) with and without closed caption data, respectively.

A. Feature selection

Similarly as in last year's experiments, we selected the set of used features for each concept detector separately using a greedy feature selection scheme [1]. For each concept, we begin with an empty set of features and keep adding the best-scoring feature as long as it improves the overall result. The optimization criterion was again the average precision at 2000 returned items with two-fold cross validation on the development set. As candidate features, we used all the visual features listed in Sections II-A and II-B, with the exception of runs F1 and F2 in which the feature selection is restricted to still-image and video features, respectively.

B. Auxiliary concepts

As part of our experiments this year, we examined the utilization of positive and negative *auxiliary* concepts. The conventional approach to automatic concept detection has been to train a detector with positive and negative examples of that concept and to do this independently of the other concepts. By contrast, in these experiments we include additional auxiliary concepts from the full LSCOM ontology to the required 39 LSCOM-Lite concept detectors, based on visual similarities and co-occurrence inter-concept relations.

The current version 1.0 of LSCOM defines 856 concepts, of which 449 have been used to annotate the TRECVID 2005 development set in a collaborative annotation process. Of these 449 concepts, 430 are marked relevant to at least one shot in the training data. However, to ensure enough training data for cross-validation, we used here only a subset of the annotated LSCOM concepts for which the proportion of relevant shots exceeded 0.001 on the shot level. This requirement reduced the number of concept models to a total of 294. The visual feature selection process described in the previous section was carried out also for the LSCOM concept models. None of the text features were used, so the LSCOM concept models correspond to the submitted concept models in the run F3.

It is clearly impractical to test every concept in a large-scale ontology as a candidate auxiliary concept in both positive and negative senses. Therefore, we used the following heuristic to select the candidate concepts. In practice, there are few if any concepts that co-occur frequently in the same shot, but are visually very different, due to the use of non-localized annotations and global

features. Still, when such concepts exists, they may be considered potentially helpful for building concept detectors as they might reveal such shots relevant to a concept that would be otherwise easily neglected. The opposite holds for concepts useful as negative auxiliaries: a visually similar but seldom co-occurring concept is likely to produce false positives.

Using these criteria, we picked out five positive and five negative LSCOM candidate concepts for each of the 39 concepts detectors. In addition, we added the concept *News Studio* as the sixth negative candidate for all detectors. We then checked the candidate concepts individually to see whether their inclusion improved the detection results, using cross-validation with the development set, and rejected those that failed to show improvement in performance. Typically this process resulted in 1–4 auxiliary concepts per detector, the majority of which were negative. All the resulting positive and negative concepts are listed in Table III.

C. Results

Instead of the standard average precision (AP) score, *inferred average precision* (InfAP) [12] was used in to evaluate this year's high-level feature extraction submissions. InfAP enables the manual judgment of only a sample of the submission pool. To generate this year's ground truth, a sample of 50 % of the pooled submissions for 20 of the 39 concepts were judged.

Figure 4a shows the mean InfAP values as an overview of our runs in the high-level feature extraction task. The highest mean InfAP score of our submitted runs was 0.0797 obtained with run 9, compared to the median of 0.0704 and best run of 0.192 over all submissions. Of our total of nine runs, the unsubmitted run 8 performed slightly better, obtaining a mean InfAP of 0.0825. The concept-wise results are illustrated in Figure 4b, which shows the best InfAP value among our submitted runs compared to the median and maximum values over all submissions. For each concept, the run which yielded the best InfAP value within our submitted runs is also shown.

As can be observed in the runs F1–F3 of Figure 4a, the combined use of image and video features results in better performance than with either of these modalities separately. This was expected based on the cross-validation experiments with the development set, so both modalities were utilized in all of the succeeding runs.

From the runs F4–F6, it can be seen that incorporating the text features shows an observable improvement in the mean InfAP values. Between the non-stemmed and stemmed ASR/MT features, and with and without the closed caption data, the differences are however minimal.

The effect of the auxiliary concepts is shown in the runs F7–F9 of Figure 4a, and in more detail in Figure 5. In the latter figure, the two submitted runs using the auxiliary concepts are compared concept-wise to corresponding runs without them, i.e. the shown comparisons are between runs F3 vs. F7 and run F4 vs. F9. Overall, we can observe a slight general improvement when utilizing the auxiliary concepts.

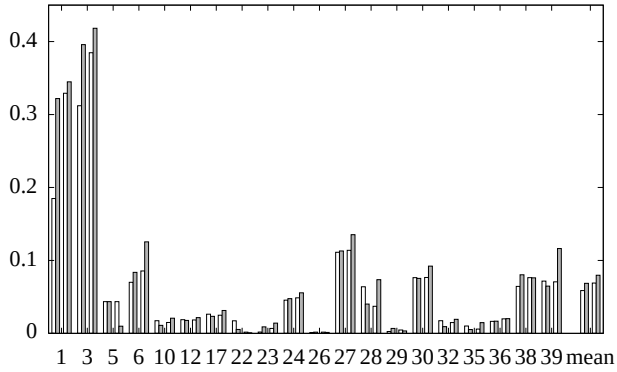


Fig. 5. A comparison of high-level feature extraction runs with (gray) and without (white) the auxiliary concepts. For each concept, the four bars correspond to InfAP values of runs F3, F7, F4, and F9.

V. Search Experiments

For the search task, we submitted five automatic runs and one interactive run. The runs are summarized in Table IV. The retrieval technique is similar to the one used in our TRECVID 2005 submission [1]. The general idea is to combine SOM-based visual features with inverse-file text features and both positive and negative concept models.

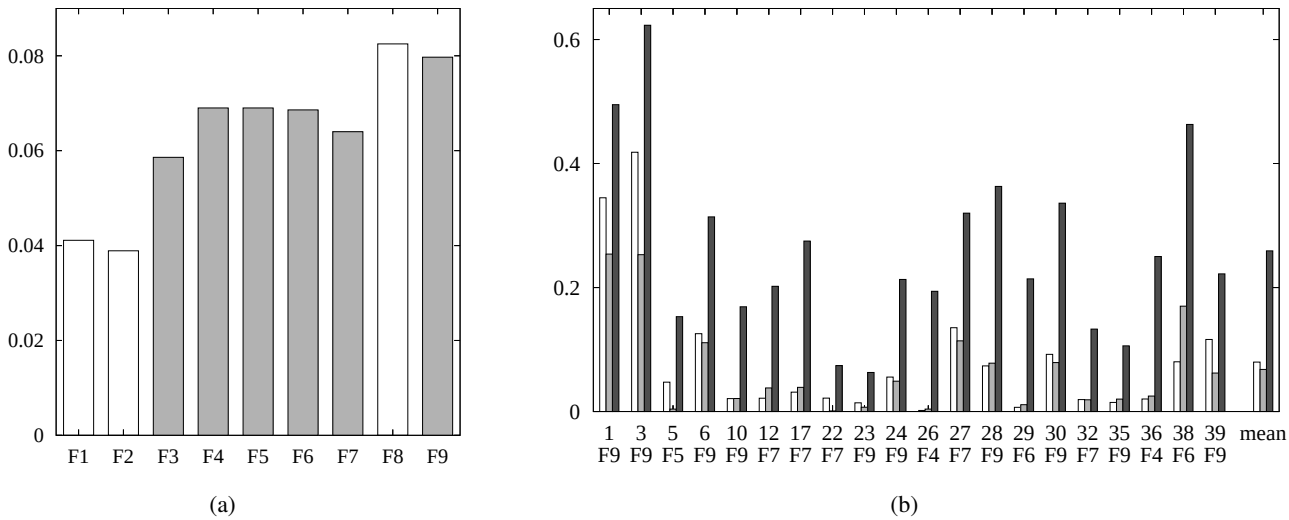
Before the search experiments, features were extracted from the provided example videos and images for each search topic. As the example videos were taken from the development set, key frames could be provided for most of them by matching the time intervals with those of the development set key frames. If no key frames were found, sufficiently "typical" frames were extracted directly from the videos themselves. After feature extraction, the best-matching map unit for each example object was located on every SOM of the corresponding modality in use and the objects were mapped to them.

A set of nine visual features (image and video) was gathered based on the feature-wise performance in the feature selection process of the high-level feature extraction task (see Section IV-A). Five video features (Color

TABLE III

THE POSITIVE AND NEGATIVE AUXILIARY CONCEPTS USED IN THE HIGH-LEVEL FEATURE EXTRACTION TASK.

LSCOM-lite concept	Positive LSCOM concepts	Negative LSCOM concepts
1: sports	-	Body Parts, Commercial Advertisement
2: entertainment	Water Tower, Amusement Park	Windows
3: weather	-	Trees, Sky, Vegetation, Overlaid Text
4: court	-	Microphones
5: office	-	News Studio
6: meeting	-	Walking Running, Scene Text, News Studio
7: studio	-	-
8: outdoor	Water Tower, Houseboat	-
9: building	Barn, Water Tower, Clocks	News Studio
10: desert	Airplane Crash	News Studio
11: vegetation	Referees	Scene Text, News Studio
12: mountain	-	News Studio
13: road	-	Group
14: sky	-	Scene Text, Office, Furniture, News Studio
15: snow	Glacier, Ski	News Studio
16: urban	Water Tower, Empire State	-
17: waterscape/waterfront	Raft	News Studio
18: crowd	Handguns	-
19: face	-	Walking, Walking Running
20: person	Studio With Anchorperson	Body Parts, Car
21: government-leader	Jacques Chirac	Scene Text, Female Person, News Studio
22: corporate-leader	-	Scene Text
23: police/security	Landing Craft	News Studio
24: military	Foxhole	Windows, News Studio
25: prisoner	-	News Studio
26: animal	-	Powerplants, Single Person Male
27: computer/tv-screen	-	Overlaid Text
28: flag us	-	News Studio
29: airplane	-	News Studio
30: car	-	Standing
31: bus	-	News Studio
32: truck	-	News Studio
33: boat/ship	-	Landscape, Sky
34: walking/running	-	-
35: people-marching	-	News Studio
36: explosion/fire	Airplane Crash	-
37: natural-disaster	Avalanche	News Studio
38: maps	-	Scene Text, Female Person
39: charts	-	Logos Full Scr., Comm. Advertisement, Overlaid Text, Person



(a) (b)

Fig. 4. (a) Mean InfAP values for our runs in the high-level feature extraction task; the submitted runs in gray. (b) The best concept-wise InfAP values from our submitted (F3-F7, F9) runs (white) compared to median (light gray) and maximum (dark gray) values over all runs.

Moments, Texture Neighborhood, Edge Histogram, Edge Co-occurrence and Average Color) and four image features (MPEG-7 Edge Histogram, MPEG-7 Color Structure, Texture Neighborhood and Color Moments) were picked as the default set of visual features in the search experiments.

Instead of the concept-specific text features used in the high-level feature extraction task, an inverse file was created using all the words in the ASR/MT output. Common stop words and very infrequent ones were not used. In all search experiments the query topic description, e.g. "Find shots with one or more emergency vehicles…", was used after some pre-processing. The text query expressions were stemmed using the Porter stemming algorithm [13], and the SMART stop list [14] for removal of common terms was applied. Furthermore, based on additional information provided with the ASR/MT output of non-English videos in TRECVID 2005, lists of proper names of persons, locations and organization could be created. This was useful for fixing misspelled names in the text query and also utilized in the retrieval process. The query terms were weighted using inverse document frequency.

A. Automatic search and concept matching

Table IV gives an overview of the experiments performed in the search task. The run numbered F1 constitutes the required baseline run using only the query texts and ASR/MT output. Run F2 uses only the default set of nine visual features, while F3 uses these combined with the textual inverted file. The run F4 uses the visual and textual features combined with concept matching based on the query text, as introduced in our experiments last year [1]. F5 was a combination of the above approaches, using either only textual search (as in F1) or only visual features combined with concept matching. The criterion was based on a division between "generic" and "specific" topics, the latter being topics with descriptions containing proper names. The specific topics were retrieved using a purely text-based approach, while the visual features and concept matching were used for the generic topics.

This year the concept matching was based on a larger set of LSCOM concepts than in 2005 (see Section IV-B). The concept names in LSCOM are descriptive, for example *Ship*, and with some pre-processing these could be passed to the WordNet [3] returning a set of synonyms. For example *Ship* would get the following get of discriminative words: ship, watercraft, vessel. If such words were present in a query description, the corresponding class models would be added as positive concepts. The presence of negative words, like a preceding "not" would negate the class model. Table V shows a list of concepts selected for the different search topics in the automatic search. Negative concepts are shown in italics.

Also new for this year was that additional *concept dependent* visual features were added based on the selected class models in experiments F4 and F5. For each LSCOM class model a set of optimal features was selected using the same feature selection algorithm that was used in the high-level feature extraction task (Section IV-A). Of course many of these features were already present in the default set of visual features, so this had an effect only on some queries.

B. Interactive search experiment

The user interface used for interactive search was a slightly modified version of the basic PicSOM user interface designed for prototyping relevance feedback based retrieval of images. As a result, the system was functional but not by any means optimal for video shot browsing and retrieval. In this experiment, textual and visual features were used in combination with string matched concepts, but without the concept-dependent features.

The system was set to always return 20 best-scoring shots. On each round, the query continues as the user assesses the returned shots and marks the ones that she considers relevant. The remaining ones are regarded as non-relevant. All previously found relevant objects are shown in the lower part of the user interface. This makes it possible for the user to remove objects from the set of relevant objects at a later stage. The user interface also supports returning to previous query rounds or back to the initial screen, where it is possible to change parameters and restart the search. When the user is finished the final 1000 best-scoring video shots are stored.

The interactive experiment was performed by seven researchers of our laboratory, five of which are not involved in our research group. They did not have any direct contact with the TRECVID 2006 test data prior to the experiment. The search sessions were limited to 15 minutes.

C. Results

The MAP scores for all our search runs are listed in Table IV. For comparison, the median and maximum values of MAP was calculated from all TRECVID participants' runs of the same type, i.e. with the same

TABLE IV

AN OVERVIEW OF SEARCH TASK RUNS. FEATURES MARKED WITH "S" ARE USED FOR "SPECIFIC TOPICS", "G" FOR "GENERAL TOPICS".

#	run id	textual	visual	conc.dep.	concepts	MAP	median	max
F1	F_A_1_OM-f1_6	•				0.0152	0.0349	0.0450
F2	F_A_2_OM-f2_5		•			0.0190	0.0243	0.0867
F3	F_A_2_OM-f3_4	•	•			0.0246	0.0243	0.0867
F4	F_B_2_OM-f4_2	•	•	•	•	0.0362	0.0390	0.0753
F5	F_B_2_OM-f5_3	S	G	G	•	0.0369	0.0390	0.0753
I1	I_B_2_OM-i_1	•	•		•	0.0510	0.1665	0.3034

TABLE V

LSCOM CONCEPTS SELECTED FOR THE SEARCH TOPICS IN AUTOMATIC SEARCH, CONCEPTS LISTED IN ITALICS ARE NEGATIVE.

Topic	LSCOM Concepts
173:emergency vehicles	Emergency Vehicles, Ground Vehicles, Police, Vehicle, Explosion Fire, Police Private Security Personnel
174:tall buildings	Building
175:leaving or entering vehicle	Ground Vehicles, Vehicle
176:escorting prisoner	Guard, Police, Soldiers, Police Private Security Personnel, Prisoner
177:demonstration or protest	Daytime Outdoor, Demonstration Or Protest, Building, People Marching
178:Dick Cheney	Head Of State, Face, Government Leader, Person
179:Saddam Hussein	Face, Person
180:in uniform and in formation	Military Personnel
181:George W. Bush	George Bush, Head Of State, Walking, Face, Government Leader, Person, Walking Running
182:soldiers or police	Armored Vehicles, Emergency Vehicles, Ground Vehicles, Police, Soldiers, Vehicle, Weapons, Military Personnel, Police Private Security Personnel
183:water with boats	Ship, Boat Ship, Waterscape Waterfront
184:seated at computer	Computers, Sitting, Computer Or Television Screens
185:reading newspaper	Newspapers
186:natural scene	Beach, Lakes, Lawn, Oceans, River, Trees, Animal, Mountain, Sky, Vegetation, *Ground Vehicles*, *Vehicle*, *Building*, *Road*
187:helicopters in flight	Flying Objects, Helicopters
188:burning with flames	Explosion Fire
189:seated group in suits and flag	Flags, Group, Sitting, Suits
190:person and books	Person
191:adult and child	Adult, Child, Person
192:kiss on the cheek	Greeting
193:smokestacks or chimneys	Smoke, Smoke Stack, Tower
194:Condoleeza Rice	Face, Person
195:soccer goalposts	Soccer
196:snow	Snow

training type (A or B) and condition value (1 or 2). For the interactive run we have calculated the median and maximum from all interactive runs regardless of type.

The results of the automatic runs indicate that the video search performance of the PicSOM system improves by augmenting a text query with visual low-level features and query-string-based matching of class models. The most notable improvement comes from the addition of appropriate concepts. Overall, the results suffer from the poor performance of the text-based baseline (run F1), which is well below the median of all submitted baseline runs. The run F5, which uses textual features for specific topics and visual features combined with concepts for generic topics, is the best run overall, although the difference in MAP values between runs F4 and F5 is quite marginal.

Figure 6 shows the topic-wise results with the runs for each topic in the same order as in Table IV. The results for topic 195 are shown in a separate graph as they are much higher than the others, going off the scale.

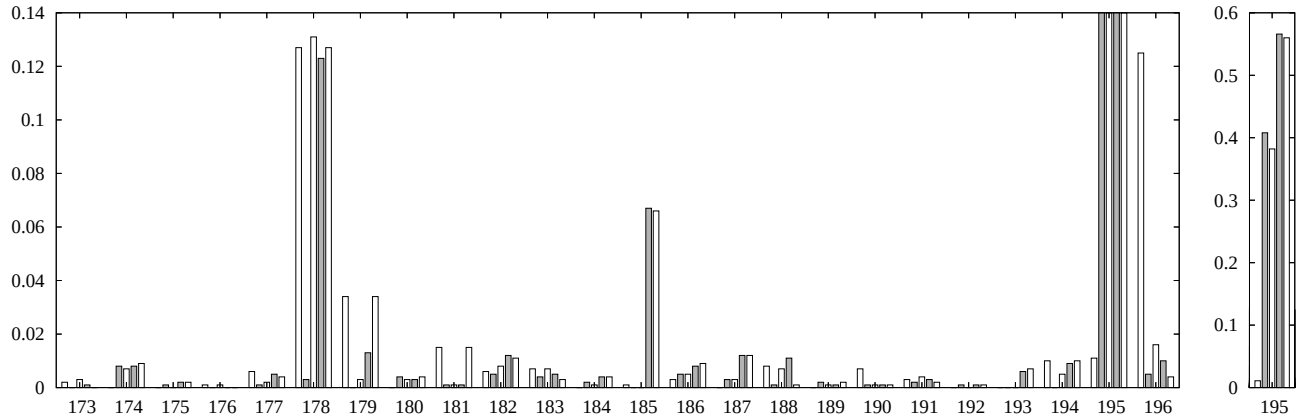


Fig. 6. Topicwise results for the automatic search runs F1 – F5.

The performance generally follows roughly the same order as with the average results, with some interesting exceptions.

For example, in topic 196 ("scenes with snow") the baseline result is dramatically better. Apparently the word "snow" was very effective, while the visual features only made the combined results much worse. One might imagine that color-based features might be good at detecting snow-filled scenes, but a visual inspection of the relevant results shows for example news stories about snow storms in cities and urban areas with buildings, emergency vehicles, and other things producing shots with many different colors. The color features tend to retrieve uniformly colored shots where the white color is very dominating, for example white graphics or over-exposed frames. On the other hand, topic 187 clearly demonstrates the power of concept matching.

In run F5, the following specific topics were identified: 178, 179, 181, and 194. If we investigate the topic-wise results we can see that our assumption regarding the specific topics was correct. In many cases the purely text-based result is indeed dramatically better than the visual results, and even in those cases where this is not true, the text-based results are still comparable.

The results of our interactive search experiment were below the median of all participants' results. This was not unexpected as the user interface was not designed or optimized for these kinds of experiments. Between topics, the results were very varying, and the greatest improvements over the automatic searches could be found in topics that are very specific, for example 189: "seated group in suits and flag".

VI. Conclusions

This was our second year participating in the TRECVID evaluations. In 2005, we demonstrated that our content-based information retrieval system PicSOM was suitable for digital video retrieval with promising results in both high-level feature extraction and search tasks. This year we wanted to try both several smaller evolutionary improvements together with some new ideas.

We participated in the shot boundary detection task for the first time this year, and had to spend a considerable amount of time to design and implement the detector. We had limited time and limited annotated shot boundary data to test the performance of our detector, but our SOM-based method seemed to work quite well. We learned a lot about the limitations of our detector from the evaluation results.

In our experiments we have observed that increasing the number of parallel detectors in the committee improves the detection precision and recall, if there is sufficiently training data to avoid overfitting. However, this also increases the computational load of the algorithm significantly. The complexity might be decreased by replacing the eleven feature extraction methods with a lesser number of SBD-optimized extraction methods. The methods utilized in our detector have been originally developed for image and video retrieval, and therefore they might not be optimal in shot boundary detection task. We should also investigate if some of the freely adjustable parameters could be fixed to some constant values to reduce the risk of overlearning.

In the high-level feature extraction and search tasks, the video objects are in PicSOM split into multiple modalities (video, audio, key frame images and ASR/MT

text data), and indexed as a hierarchical structure with several parallel SOMs using relevance propagation between the multiple modalities. Class models created from the common annotation of the development data were already used successfully in 2005.

In the high-level feature extraction task our methodology was very similar to the one used in 2005. However, we now experimented with using auxiliary concepts from the full LSCOM ontology as positive and negative class models selected automatically for each high-level feature detector. This technique improved our extraction results.

In the search task we submitted a set of automatic runs and one interactive run. In addition to the new class models based on LSCOM concepts we employed a smart feature selection algorithm that optimized the concept-wise feature sets. In most cases adding visual features and concepts to the baseline textual features significantly improved the retrieval performance. The overall results clearly suffered from the poor performance of the text baseline. Still, as we had already previously observed, using only textual features can perform better in some cases. To this end we devised a method for detecting named entities, or "specific" topics, in the textual search topic descriptions. Such topics were processed using only textual features, which improved the search performance.

ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland in the projects *Neural methods in information retrieval based on automatic content analysis and relevance feedback* and *Finnish Centre of Excellence in Adaptive Informatics Research*.

This work was supported by The Irish Research Council for Science, Engineering and Technology.

REFERENCES

[1] Markus Koskela, Jorma Laaksonen, Mats Sjöberg, and Hannes Muurinen. PicSOM experiments in TRECVID 2005. In *Proceedings of the TRECVID 2005 Workshop*, pages 262–270, Gaithersburg, MD, USA, November 2005. Available online at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[2] DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia. LSCOM lexicon definitions and annotations version 1.0. Technical Report #217-2006-3, Columbia University, March 2006.

[3] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[4] Jorma Laaksonen, Markus Koskela, and Erkki Oja. PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. *IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing*, 13(4):841–853, July 2002.

[5] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer-Verlag, Berlin, third edition, 2001.

[6] Mats Sjöberg and Jorma Laaksonen. Content-based retrieval of web pages and other hierarchical objects with Self-Organizing Maps. In *Proceedings of 15th International Conference on Artificial Neural Networks (ICANN 2005)*, pages 841–846, Warsaw, Poland, September 2005. Available online at http://dx.doi.org/10.1007/11550907_133.

[7] ISO/IEC. Information technology - Multimedia content description interface - Part 3: Visual, 2002. 15938-3:2002(E).

[8] Markus Koskela, Jorma Laaksonen, and Erkki Oja. Use of image subset features in image retrieval with self-organizing maps. In *Proceedings of 3rd International Conference on Image and Video Retrieval (CIVR 2004)*, pages 508–516, Dublin, Ireland, July 2004.

[9] Hannes Muurinen and Jorma Laaksonen. Video segmentation and shot boundary detection using self-organizing maps. In *Proceedings of 15th Scandinavian Conference on Image Analysis (SCIA 2007)*, 2007. Submitted.

[10] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann. A light scale concept ontology for multimedia understanding for TRECVID 2005. Technical report, IBM, 2005.

[11] Milind Naphade, John R. Smith, Jelena Tešić, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[12] Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of 15th International Conference on Information and Knowledge Management (CIKM 2006)*, Arlington, VA, USA, November 2006.

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

[14] Gerard Salton, editor. *The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall, 1971.