PicSOM Experiments in TRECVID 2009

Mats Sjöberg, Ville Viitaniemi, Markus Koskela, Jorma Laaksonen
Adaptive Informatics Research Centre, Department of Information and Computer Science
Helsinki University of Technology (TKK), Finland

Abstract

Our experiments in TRECVID 2009 include participation in the high-level feature extraction and automatic search tasks.

In the high-level feature extraction task, we used a feature fusion-based general system architecture utilizing a large number of SVM detectors, followed by a post-processing stage utilizing the concepts' temporal and inter-concept co-occurrences. We submitted the following six runs:

- `PicSOM.base`: Baseline run using our SOM-based HLF detection method
- `PicSOM.A-ngram`: Baseline SVM-based run using HLF-wise geometric mean fusion and temporal n-gram post-processing
- `PicSOM.B-ngram`: As previous, but includes also early fusion, multi-fold SFBS fusion, and more elaborate SVM training
- `PicSOM.E-ngram`: As previous, but includes two-stage fusion utilizing cross-concept co-occurrence
- `PicSOM.spec-ngram`: A run where the used method was selected for each HLF separately using cross-validation
- `PicSOM.spec-any`: As previous, but the post-processing used also clustering-based inter-concept co-occurrence analysis

The results show that feature fusion can consistently outperform all single features, multi-fold SFBS performed best of the tested fusion methods, and that temporal n-gram analysis is beneficial. Early fusion, and post-processing based on inter-concept co-occurrences did not improve the performance.

In the search task, we concentrated on the fully-automatic runs and standard search task. We combined ASR/MT text search and concept-based retrieval. If none of the concept models could be matched with the query, we used content-based retrieval based on the video and image examples instead. We submitted the following ten fully-automatic runs:

- `F_A_N_PicSOM_1_10`: text search baseline
- `F_A_N_PicSOM_2_9`: visual baseline
- `F_A_N_PicSOM_3_8`: own concepts
- `F_A_N_PicSOM_4_7`: own concepts + text search
- `F_A_N_PicSOM_5_6`: donated concepts
- `F_A_N_PicSOM_6_5`: donated concepts + text search
- `F_A_N_PicSOM_7_4`: own + donated concepts
- `F_A_N_PicSOM_8_3`: own + donated concepts + text search
- `F_A_N_PicSOM_9_2`: own + donated (dupl.) concepts
- `F_A_N_PicSOM_10_1`: own + donated (dupl.) concepts + text search

In the above list, "own" concepts refer to our own HLF detectors and "donated" concepts consist of MediaMill (MM) concepts + CU-VIREO374 concepts. In other than the last two runs, CU-VIREO374 are only used for words for which no MediaMill concept could be matched. The results show again that concept-based retrieval performed better than content-based search alone. Text search made a small improvement in combination with other modalities, but performed really badly on its own. Concept-selection was done both with word-matching and example-based matching, i.e. selecting concepts based on how well they would fit our own concept models.

I. Introduction

In this paper, we describe our experiments for the TRECVID 2009 [1] evaluations. We participated only in the high-level feature (HLF) extraction and automatic search tasks. This year, we essentially substituted our Self-Organizing Map (SOM) based analysis framework (e.g. [2], [3]) with SVM-based classifiers for the high-level features. In automatic search, we combined text search with HLF classifiers trained by us and with HLF classifiers donated by the MediaMill and CU-VIREO teams. SOM-based retrieval using the PicSOM system [4] was used in HLF detection as a baseline system and in search when no HLF detectors were available.

The rest of this notebook paper is organized as follows. The low-level features used in both tasks are briefly described in Section II. Our experiments for the HLF extraction and search tasks are described in Sections III and IV, respectively. The overall conclusions are presented in Section V.

II. Low-level features

We extracted in total 3 video and 18 still image (keyframe) features. The keyframes were extracted from the video shots in the master shot reference [5] using a heuristic algorithm (see [2]). Separate SOMs of size 256×256 map units were then trained for each of the video and image features.

The used features are briefly described in Sections II-A to II-C.

A. Image features

For the video keyframes and image examples we used a set of standard image features. Our extracted features include five MPEG-7 descriptors implemented in the XM reference software, our own implementations of four of the MPEG-7 descriptors, and seven other image features: *Average Color*, *Color Moments*, *Texture Neighborhood*, *Edge Histogram*, *Edge*

Co-occurrence, *Edge Fourier*, and *SIFT (ip)*. See [2] for details of these features.

In addition we extracted ColorSIFT features [6] using the opponent color space and with two different sampling strategies: the Harris-Laplace salient point detector (*Color SIFT (ip)*) and dense sampling (*Color SIFT (dense)*). The codebooks for both variations were generated by first taking a random sample of 100 keyframes and calculating the features for all of their sampled points. The resulting vectors were partitioned into 1000 clusters using k-means. The cluster centroids were then selected as the codebook vectors.

B. Video features

For the video shots we used temporal extensions of three of the calculated still-image features: one of our own MPEG-7 implementation of *Color Layout* and two variations of *Edge Histogram*. See [2] for details.

C. Text features

The Dutch automatic speech recognition (ASR) output [7] was machine-translated (MT) to English. We used only the English documents on the shot level using temporal smoothing (see Section IV-B). Similarly as in previous years' experiments, we indexed the ASR/MT output using the Apache Lucene text search engine. The Snowball stemmer included in Lucene was used with its default stop word list.

III. High-level Feature Extraction

This year we addressed the high-level feature (HLF) extraction task with a well-established feature fusion-based general system architecture: dozens of supervised detectors were trained for each HLF, based on different shot-level image and video features detailed in Section II. The probabilistic detector outcomes were then fused. Finally, the fused detection scores were re-adjusted based on the detector outcomes for temporally neighboring video shots.

This system architecture has been used rather successfully in various multimedia content analysis tasks, including TRECVIDs of previous years. This year we wanted to examine whether the architecture still could provide reasonable HLF detection performance if some of the system components were updated. We also studied the effect of applying different alternative techniques in various stages of the detection system.

An identical procedure was used for detecting all the HLFs. As the concept-wise ground-truth for the supervised detectors we used the annotations gathered by the organized collaborative annotation effort [8]. All our submitted runs were of type A. To validate and compare different alternative techniques prior to submission we performed validation experiments with a 2:1 split of the development data.

A. Components of the detection system

Our detector system contains many components that could be implemented in different ways. In our experiments, we studied some of these alternatives. Starting from the shot-level feature extraction, our baseline alternative was to use all the 21 extracted image and video features of Sections II-A and II-B. Additionally, we examined the option of augmenting the feature set with early fusion, i.e. by concatenating some of the feature vectors together after normalizing their components.

In our system, a number of feature and HLF specific supervised detectors is trained based on the extracted features. Previously, we have used a Self-Organizing Map (SOM) based algorithm as this probabilistic supervised classifier component. This year we replaced SOMs with Support Vector Machines (SVM), motivated by large accuracy improvements we have been observing in other classification tasks. We evaluated, however, also one SOM-based run as a baseline.

The SVM implementation we used was an adaptation of the C-SVC implementation of LIBSVM software library [9]. The SVM parameters were selected with an approximate 10-fold cross-validation search procedure that consisted of a heuristic line search to identify a promising parameter region, followed by a grid search in that region. The RBF kernel was used for all the visual features. In addition, we also used χ^2 kernel for some of the visual features, resulting somewhat more detector outcomes to be fused. For SVM training we employed two strategies: faster training with rather radical sampling (at most 5000 training shots out of 36000 retained, including all the positive examples), and more elaborate training with more conservative sampling (12000 training shots retained).

We tried several alternative algorithms for fusing the detector outcomes. As a baseline approach, we considered the geometric mean of all the outcomes. Besides this unsupervised fusion approach, we tried several supervised fusion methods that require the detector outcomes also for the training set. These were obtained with 10-fold cross-validation. One supervised technique was SVM-based fusion employing RBF kernels, another Bayesian Binary Regression (BBR) [10]. The other alternatives we tried were variations of the scheme where the basic fusion mechanism is still the geometric mean, but the mean is calculated only of a subset of the detector outcomes, selected by a sequential forward-backward search (SFBS). In addition to basic SFBS, we tried the idea of reserving part of the training set for validation and early-stopping the search based on the performance in this validation set. We also tried partitioning the training set into six folds. The SFBS algorithm was run six times, each time leaving one fold outside. The fusion outcome was the geometric mean of the geometric means.

Besides the fusion algorithm, we also experimented with the selection of the set of detectors that were fused for each HLF. Our basic alternative was to fuse the outcomes of the detectors that were trained for detecting this particular HLF. We also tried to exploit inter-concept co-occurrence by including detectors trained for all the other HLFs in the fusion. This idea was implemented as a two-stage fusion scheme where the detectors for each HLF were first fused separately. In the second stage, the HLF-wise fusion was repeated otherwise in the same way, but the set of detector outcomes to be fused was augmented with fused detection outcomes for all the other HLFs and their temporally smoothed versions.

TABLE I
AN OVERVIEW OF THE SVM-BASED RUNS IN THE HIGH-LEVEL FEATURE EXTRACTION TASK. SEE TEXT FOR DETAILS.

#	run id PicSOM.+	fusion early	fusion cross-c.	post-proc. n-gram	post-proc. co-occ.	MIAP
4	base	-	-	-	-	0.039
1	A-ngram			•		0.144
2	B-ngram	•		•		0.147
3	E-ngram	•	•	•		0.138
6	spec-ngram	○	○	•		0.151
5	spec-any	○	○	•	○	0.143
	additional run A1	•	•	•		0.136
	additional run A2	•	•	•		0.136
	additional run A3	•	•	•		0.142
	additional run A4	•		•		0.136

For temporal post-processing of the fusion outcomes, we employed the techniques detailed in [11]. Those techniques consist of an intra-concept n-gram smoothing technique, and inter-concept techniques based on clustering. In our binary n-gram technique the cross-validated fusion outcomes of the training set are used for estimating a sigmoidal mapping model from the fusion outcomes to probabilities of observing HLFs. This estimation is performed separately for each different temporal n-gram neighborhood. The models are applied to test data by estimating the likelihood of each n-gram on basis of the fused detector outcomes. The probabilities given by the mappings conditional to different n-grams are averaged, weighted by the estimated likelihoods.

Previously, the n-gram technique has been far more useful than the inter-concept techniques that have sometimes been even harmful. We thus considered the n-grams as our primary temporal technique and evaluated the inter-concept techniques just for reference. We selected the optimal n-gram order for each HLF separately based on the 2:1 validation experiment. We also evaluated the approach where the n-gram order was selected jointly for all HLFs. We also briefly investigated the idea of replacing the models specific for each different n-gram with models conditional to the count of occurrences of the HLF within a temporal neighborhood.

B. The submitted and additional runs

This section details the six submitted HLF runs as well as some additional runs that were used as components in the submitted runs. Table I shows an overview of the runs. Rows of the table correspond to the runs. For the SVM-based runs (i.e. all other runs but `PicSOM.base`), the columns refer to whether early fusion, cross-concept late fusion, the temporal co-occurrence techniques, and temporal inter-concept co-occurrence techniques were used. The "○"s in the table denote HLF-wise selection of that attribute in the corresponding run. The rightmost column lists the corresponding mean inferred average precision (MIAP) [12] values. In addition, Figure 1 shows the IAP results of our submitted runs for each evaluated concept.

Because of the large number of combinations of variable components in our system, the submitted runs were not enough to systematically compare the alternatives. We thus additionally evaluated dozens of runs ourselves after the benchmark in order to be able to draw some specific conclusions on the components. These runs are not addressed in this section.

The run `PicSOM.base` utilizes the SOM-based classifier algorithm, as used in our last year's experiments [3]. Exhaustive feature selection was performed with twofold cross-validation on the development set for each concept separately. The features were selected from a set of nine image and video features that performed best last year plus the two new ColorSIFT-based ones (a subset of the features presented in Section II).

The run `PicSOM.A-ngram` is based on all the extracted visual features, for some of which also χ^2 SVM-detector was trained. The faster training scheme was used for the SVM-detectors. Fusion was performed separately for each HLF by taking the geometric mean of all the detector outcomes. Temporal post-processing was performed using the n-gram technique with HLF-wise n-gram order selection based on the validation experiment.

For the run `PicSOM.B-ngram` the feature set of `PicSOM.A-ngram` was augmented with 11 feature combinations resulting from early fusion of some of the original features. Two sets of detector outcomes were included in fusion: those resulting both from the faster and from the more elaborate SVM training schemes. Fusion was performed separately for each HLF with early-stopping multifold SFBS. The temporal post-processing was similar as in `PicSOM.A-ngram`.

The run `PicSOM.E-ngram` utilized the two-stage fusion mechanism to exploit cross-concept co-occurrence. The first fusion stage fused the full set of SVMs, just as in run `PicSOM.B-ngram`. For each HLF in turn, the first stage fusion results for all HLFs were combined with limited set of those HLF-specific SVM detector outcomes that were trained with the faster SVM training scheme. The fusion algorithm and temporal post-processing were similar as in `PicSOM.B-ngram`.

The unsubmitted run A1 was otherwise similar as `PicSOM.E-ngram` but both fusion stages utilized the full set of SVM detectors obtained with both faster and more elaborate training.

In the unsubmitted run A2 the starting point for both fusion stages utilized the restricted set of SVM detectors obtained with only the faster training scheme. Otherwise the run was just as `PicSOM.E-ngram`.

In the unsubmitted run A3 the first fusion stage used the restricted set of SVM detectors, whereas the second fusion stage used the full set of SVM detectors. Otherwise the run was just as `PicSOM.E-ngram`.

The unsubmitted run A4 was otherwise similar as `PicSOM.B-ngram`, but only the restricted set of SVM detectors obtained with the faster training scheme was used as starting point for fusion.

The run `PicSOM.spec-ngram` was synthesized from the above described runs (excluding `PicSOM.base`) by selecting

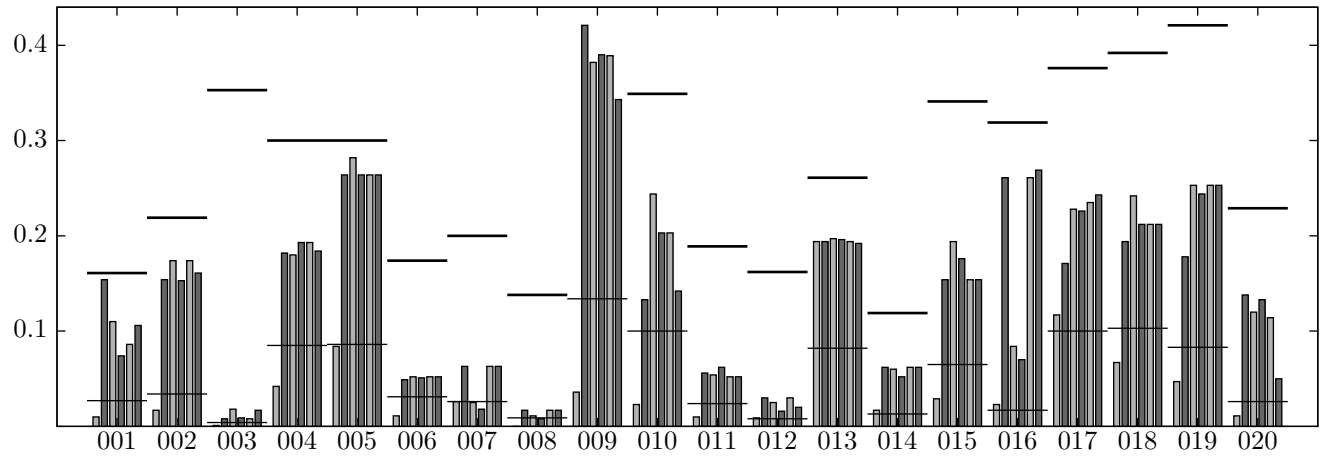

Fig. 1. The HLF-wise IAP results of our submitted runs for each evaluated HLF. The order of the runs is as in Table I. (i.e. the leftmost bar corresponds to `PicSOM.base`, etc.) The median and maximum values over all submissions are illustrated as horizontal lines. The maximum IAP score of concept 009, i.e. 0.566, is not within shown range.

for each HLF the technique that worked best among the seven alternatives for that HLF in the 2:1 validation experiment. For five HLFs the selected run was not among the three submitted ones.

The run `PicSOM.spec-any` was formed similarly, but the pool of seven runs among which the best was selected was increased to 14 by extending the temporal n-gram post-processing with the cross-concept co-occurrence techniques of [11].

C. Conclusions from the HLF detection experiments

In this section we present our conclusions from the high-level feature detection experiments. Partially these conclusions are based on the evaluation of the submitted runs. More importantly, we evaluated dozens of additional runs in order to be able to compare different component techniques of the detection system somewhat systematically. In most of the results, the fluctuation in IAP results for individual HLF was observed to be so strong that we felt we could not reliably distinguish different algorithmic techniques. Thus we settle for drawing conclusions on average only, i.e. on basis of MIAP which we consider more reliable. However, there probably are genuine differences between the HLFs—one technique might really be suitable for detecting one HLF and some other technique other HLFs. These differences will be missed by the average case analysis.

The experiments were able to to clearly confirm our prior understanding that SVM detectors are vastly more accurate than SOM-based detectors. We were also able to confirm that fusion of information given by numerous features clearly outperforms good individual features, even with a somewhat simplistic fusion algorithm. The best individual feature (*Color SIFT (dense)*) produced MIAP of 0.099, whereas the HLF-specific fusion (with multi-fold SFBS fusion and no temporal post-processing) runs produced MIAPs ranging from 0.130 to 0.139. A manifestation of the HLF-wise fluctuation is that for three HLFs, at least one of the individual features was better than the best one of the submitted fusion runs (interestingly, neither ColorSIFT feature was the best individual feature in any of these cases).

TABLE II
BEST INDIVIDUAL FEATURES IN THE HLF DETECTION TASK

Feature	SVM training	kernel	MIAP
Color SIFT (dense)	elaborate	χ^2	0.0988
Color SIFT (dense)	fast	χ^2	0.0916
Color SIFT (dense)	elaborate	RBF	0.0711
Color SIFT (ip)	elaborate	RBF	0.0657
Color SIFT (ip)	fast	RBF	0.0620
SIFT (ip)	elaborate	χ^2	0.0832
SIFT (ip)	fast	χ^2	0.0813
SIFT (ip)	elaborate	RBF	0.0769
Edge Histogram (video)	elaborate	χ^2	0.0625
Color Moments (image)	elaborate	RBF	0.0438
Edge Histogram (image)	elaborate	χ^2	0.0403

Table II lists a selection of the most accurate individual feature/SVM combinations in terms of MIAP (some ColorSIFT and SIFT variants left out). On average, the more elaborate SVM training produced somewhat more accurate detectors than faster training for 80 % of the visual features. For histogram type features, χ^2 SVM kernels produced clearly more accurate results than the RBF kernel.

This time, augmenting the feature set with early fusion, did not improve performance of the whole system. In some previous cases early fusion has been beneficial. On the level of single SVM detectors, combined features surely resulted in more accurate detectors than their individual constituent features (best early fusion combination had MIAP 0.0601). One explanation is that the best individual features were not involved in the early fusion.

Of the late fusion mechanisms, SVM and BBR based fusion

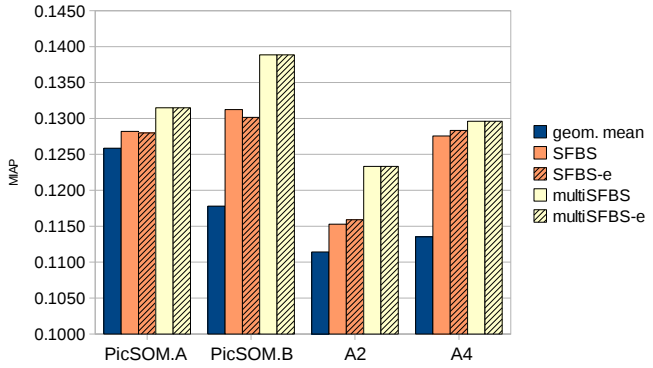

Fig. 2. Comparison of algoritms for selecting detectors for geometric mean fusion for four diffent sets of detectors. The bars with diagonal hatching correspond to algorithms with early stopping.

were significantly outperformed by geometric mean based fusion in the 2:1 validation experiment. Moreover, the SVM and BBR fusion mechanisms were computationally much more costly. Consequently the mechanisms were not used in the actual HLF detection runs. Figure 2 compares different geometric mean based fusion algorithms for the late fusion tasks of the runs `PicSOM.{A,B}-ngram` and the additional runs A2 and A4 (cf. Section III-B). We see that geometric mean of all detectors (leftmost bar) is always inferior to methods where set of detectors is selected with sequential forward/backward search (SFBS). This has not always been the case as SFBS easily overfits to the training data. This tendency might get pronounced when the training data is not actually similarly distributed as the test data, as probably is the case with the TRECVID HLF develeopment and test data. The figure also shows that multifold-SFBS performs better than basic SFBS. Early stopping seems to have no essential effect on the average performance. It may, however, increase variance of the results. Early stopping is not a proper way to regularise SFBS.

On average, we could not exploit cross-concept co-occurrence for our advantage when selecting the set of SVM detectors that were fused, although some of the runs utilizing cross-concept fusion were chosen when forming the composite runs `PicSOM.spec-ngram` and `PicSOM.spec-any`. However, here the experiment was not controlled in the sense that the property of runs using cross-concept fusion was mixed with other properties of the runs, such as the SVM training strategy. On the other hand, although the cross-concept runs seemed attractive for these particular HLFs in the validation experiment, no clear advantage for the cross-concept runs over the other runs for the same HLFs was repeated in the test set.

Should one expect to benefit from instantaneous cross-concept co-occurrence? In the limiting case with enough training data and flexible enough learning algorithms to accurately approximate the conditional probabilities involved, the answer is negative as long as the training data is fully annotated and all the concepts are detected based on the same visual features. The situation is the detections for different concepts come from independent sources. In the practical case with limited training data and learning algorithms, there probably is no general guarantee to one direction or another.

For non-instantaneous cross-concept occurrence the situation is somewhat different if the dependency occurs outside the temporal window whose shot features are used for the concept detection. For example, in our case the temporal window is just one shot, i.e. the concepts are detected based on the features of a single shot only. In principle, we could therefore be able to benefit from non-instantaneous cross-concept co-occurrence. For example we could be able to learn rules of type

```
If Concept1 at time n,
then Concept2 at time n+1 with p=0.8
```

through the temporal smoothing included in the two-stage cross-concept fusion algorithm.

In the experiments, however, we did not observe benefit from the use of the two-stage algorithm. There are several possible explanations. It is possible that there simply are no such non-instantaneous cross-concept dependencies in the TRECVID HLF annotations that could be exploited. On the other hand, we might have not been able to learn such dependencies reliably from the training data due to limited amount of independent training examples. Although the 2009 development data for the HLF task consists of over 36000 shots, there are not that many independent temporal dependency pattern examples. The number might be closer to the number of videos, which is just 219 for the development data. It is also questionable whether the temporal dependency patterns are similar enough in the training and test data in order to be exploited in a straightforward fashion. One more likely explanation is the crudeness of our fusion algorithm. In practice, it assumes all the input variables (detectors) to be symmetric in terms of the way they affect the outcome. For example, the detectors' output can not be weighted or have a negative influence on the fusion outcome, not to mention more complex non-multiplicative dependencies.

Figure 3 shows the effect of temporal post-processing for the systems of the submitted SVM runs. The darker bars indicate the runs that were submitted. The lighter bars correspond to additional runs we evaluated ourselves. We can observe that the n-gram post-processing of the submitted runs improves MIAP markedly over the baseline with no post-processing. For the submitted runs, the n-gram order was selected for each HLF separately based on the 2:1 validation experiment. This was not a good choice for strategy. We would have done better if we had chosen the n-gram order globally for each run and used it for all concepts. This may indicate that our validation setup was inadequate: maybe a better use of the training data could have been obtained with a more elaborate cross-validation setup. On the other hand, there might simply be too few independent temporal patterns in the training data, just as argued before. Similarly, the explanation of different temporal patterns in training and test data might be valid also here. This is supported by the anecdotal piece

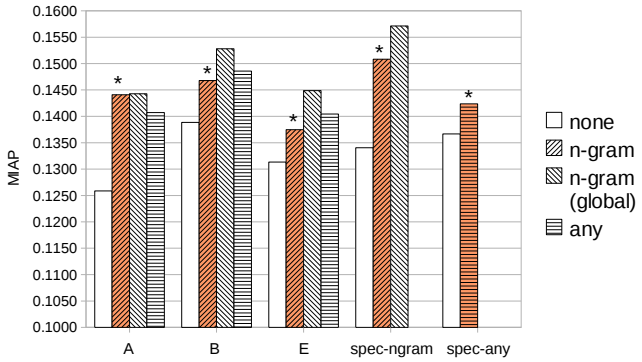

Fig. 3. The effect on temporal post-processing in several different runs. The bars marked with an asterisk correspond to submitted runs.

of evidence that in many cases the validation experiment failed to indicate the usefulness of n-grams for concept 1016 (*people dancing*), although in the test data n-grams hugely improved the detection accuracy (IAP from 0.107 to 0.312 for `PicSOM.A-ngram`).

Afterwards, we tried replacing the n-gram description of the temporal neighbourhoods with simpler descriptions, the counts of occurrences of a HLF in the same n-step neighbourhoods. The resulting detection accuracy was slightly better but still almost identical for all HLFs in the 2:1 validation setup.

After the submission time, we tried forming rank-based fusions of a few different combinations of the runs `PicSOM.{A,B,E}-ngram` and A1–A4. The fusions runs were better than any of the individual runs involved, but slightly worse than the speculative selections in the style on run `PicSOM.spec-ngram`. This shows that the 2:1 validation experiment was able to reveal genuine differences between different HLFs afterall, not all the differences were statistical fluctuations. Fusion based on detector outputs would probably have worked better than fusion based on rank.

IV. Automatic Search

For the search task, we submitted ten automatic runs summarized in Table III. All runs were submitted to the standard search task and were trained only on common TRECVID development data, thus qualifying them as type A runs.

A. The submitted and additional runs

Run 1 is a baseline run with only text-based search using the Lucene ASR/MT text index and the topicwise textual queries. Run 2 is the visual baseline, which uses content-based retrieval based on the SOM feature indices. This is the same baseline system that we have used in previous years (see e.g. [2], [3]), except for a different set of features used this year, viz. *Color SIFT (ip)*, *Color SIFT (dense)* and *SIFT (ip)*. Run 2 was used in runs 3, 5, 7 and 9 for topics for which no matching concepts could be found. The run 2b is a combination of visual content-based retrieval and text-based search. This run was not submitted as such, but was used in runs 4, 6, 8 and 10 for topics for which no concepts could be matched.

The other runs all use concept matching, as using HLF detectors has been observed to consistently surpass the retrieval performance of content-based retrieval in previous TRECVID experiments [3], [13]. In these experiments, we apply both text-based and example-based concept matching for different sets of HLF detectors. However, if no concepts can be found for a particular search task we have used the baseline visual (run 2) or visual + text (run 2b) as appropriate. This secondary use of visual examples is shown with a "$\circ$" in Table III. The concept matching procedure is described in more detail in Section IV-C.

Run 3 uses a set of 30 of our own concept detectors (*own set*) trained as described in Section III. We used the `B-ngram` method, which had the highest MIAP score for HLF detection in the development set. Of these, 20 were the ones submitted in this year's HLF detection task, and the remaining 10 were the concepts used in 2008 but not this year.

These concepts were selected in two ways: by matching with the provided visual examples of the search topics, and by automatic detection of relevant words in the search queries. To illustrate the effect of both selection methods, we created two additional runs: Run 3b was performed using only word-based selection (marked by 'W' in Table III), and 3c with only visual example based selection (marked with 'E' in the table).

The runs 5–10 use also the *donated set* of concepts, which contains concepts from the set of 64 concepts shared by MediaMill (MM) this year [14] and from the 374 CU-VIREO374 [15] (CU) concept detectors.

Runs 5–8 use the primary matching scheme, where initially the MM concepts are matched to the textual query and CU concepts are then matched for such words for which no MM concepts were found.

The additional run 8b is based on run 8 but with some modifications in concept selection (marked with "$\bullet\star$" in Table III). First, we included those concepts from the donated set that had been selected based on the visual examples in our own set. That is, if a specific concept was selected purely by visual examples from our own set of concepts, e.g. *cityscape* for topic 0269, the corresponding concept from MediaMill would be added in this run. The run also includes some retrospective changes in the concept selection, as some obvious errors were corrected manually. We also assigned to all duplicated concepts a weight of 0.5, and gave the concept *person* a lower weight to reflect that it is a very generic concept.

In runs 9 and 10, we include all matched donated concepts. This means that for many words there will be duplicates, i.e. concepts matched from both the MM and the CU sets. This selection method is illustrated with "$\bullet+$" in Table III.

B. Text search

For text-based search, the topic-wise English queries were analyzed using the Stanford part-of-speech (POS) tagger [16]. The nouns, verbs, and adjectives of each query were used as the text search queries, expanded with synonyms using the WordNet [17] package included in the Lucene search engine.

TABLE III
AN OVERVIEW OF THE SEARCH TASK RUNS. SEE TEXT FOR DETAILS.

#	run id	text	visual	concepts own	concepts donated	MAP
1	F_A_N_PicSOM_1_10	•				0.0042
2	F_A_N_PicSOM_2_9		•			0.0093
2b	additional run	•	•			0.0095
3	F_A_N_PicSOM_3_8		○	•		0.0395
3b	additional run		○	W		0.0295
3c	additional run		○	E		0.0345
4	F_A_N_PicSOM_4_7	•	○	•		0.0400
5	F_A_N_PicSOM_5_6		○		•	0.0849
6	F_A_N_PicSOM_6_5	•	○		•	0.0854
7	F_A_N_PicSOM_7_4		○	•	•	0.0910
8	F_A_N_PicSOM_8_3	•	○	•	•	**0.0913**
8b	additional run	•	○	•	•⋆	0.1030
9	F_A_N_PicSOM_9_2		○	•	•+	0.0729
10	F_A_N_PicSOM_10_1	•	○	•	•+	0.0745

The ASR/MT documents were used on the shot level. The shot-level retrieval results were spread to the temporally neighboring shots using a triangular kernel of five shots in width.

C. Semantic concept matching

In runs 3–10, the search topics were matched with the semantic concepts using word matching. For each concept, a word list was generated by taking the concept name itself as the initial word or words and expanding with WordNet synonyms. These lists were then cleaned up by hand (without knowledge of the particular search topics). For example, words with too broad meaning were removed, e.g. the concept *people dancing* was set to be activated for the word "dance" but not for "people" appearing in the textual query.

In addition, our own concepts were also matched by using the visual image and video examples given with each search query. This was done using a similar method as the HLF run `PicSOM.A-ngram` (excluding the temporal post-processing) so that for each example image or video we get a matching score for each concept. The scores were fitted to a logistic sigmoid model to be able to map them as probabilities. These were then summed up over all examples for each search query so that we end up with a matching score for each concept to that particular query. Then for each concept we select those search topics for which the score is more than six times larger than the median score for that concept. This heuristic limit was selected by two rules of thumb: we wanted the number of concepts selected per task to be on average close to one. The second rule of thumb was that there shouldn't ever be more than three concepts selected for a single topic, since from experience a small number of concepts per topic seem to work the best.

Table IV summarizes the selected concepts. The first column shows the topic number, the second our own concepts selected by visual example matching. The third column shows those of our own concepts selected by word matching. Roughly 40% of the concept selections were added due to the visual matching only, i.e. they would not have been found using only word-based selection.

The forth column in Table IV shows the concepts selected with the primary matching scheme from the donated set (i.e. as used in runs 5–8), using only word-based selection. These are mostly from MediaMill, those from CU-VIREO374 are prefixed with CU-. Due to an error in our word-based selection the concept CU-Talking was inadvertently selected for the search topic 0271. Also, in the topic 0276 the concept *person walking or running* was inappropriately selected for a query about dogs running. For topics 0280 and 0285 no concepts could be selected. For these runs we used the visual or visual + text baselines as appropriate.

D. Results

As can be seen from the overall results in Table III, the best performing submitted run used the combination of our own concepts and the donated set of concepts (run 8). In this run our own concept detectors were selected by both word-matching and based on the visual examples. The additional concepts were selected only based on word-matching.

The unsubmitted run 8b obtained a rather notable improvement, but it includes some additional HLF detectors from the donated set and some additional weighting done in retrospect, and some manual corrections. Therefore, this run cannot be directly compared to the other automatic runs. However, with a future improved automatic word-based selection this result might be achieved.

The use of visual-example based matching of concepts was very beneficial, especially using only our own concepts. This is illustrated by the success of run 3c compared to 3b. Using both ways of selecting concepts together gives an even better result (run 3).

We can also observe that although the text search performed quite badly (run 1) on its own, it still always improved the results in combination with the other methods.

The topic-wise results are summarized in Figure 4. For a more concise visualization we have omitted the concept-based runs without text-search since the text-search consistently improved the runs by a small amount. We have also omitted runs 9–10 since they performed relatively worse and don't provide much useful comparison. The bars are thus in the following order: text-baseline (run 1), visual-baseline (run 2, darker bar), own set of concepts (run 4), donated set of concepts (run 6, dark), own + donated concepts (run 8) and finally unsubmitted run 8b (dark). In run 8b, the great improvement in topic 0276 is due to removing the concept *people walking or running* which was incorrectly selected by the automated selection.

V. CONCLUSIONS

In the experiments reported in this paper, we utilized the SVM-based concept classification approach for HLF detection and used these detectors in automatic search.

In HLF detection, we applied a fusion-based general system architecture, which considers a large number of potential SVM

TABLE IV
CONCEPT SELECTION

query	own concepts (by examples)	own concepts (by words)	donated set
0269	Cityscape, Driver, Street	Street	Road, Street, CU-Vehicle
0270	Demonstration Or Protest	Demonstration Or Protest	Crowd, Demonstration Or Protest, Person, Outdoor
0271	Cityscape, Bridge		(CU-Talking), Building
0272	Telephone	Telephone	Person, Talking, Telephone
0273	Hand	Hand	Hand, Charts
0274	Two People	Chair	Two People, Person, Chair
0275	Doorway		Person, Walking Or Running
0276			Dog, Walking Or Running
0277			Person, Talking, CU-Microphones
0278	Doorway, Street	Doorway	Building, Doorway
0279		Hand	Person, CU-Handshaking, Hand
0280			
0281	Playing Musical Instrument, Singing	Playing Musical Instrument, Singing	Two People, Person, Singing, Playing Musical Instrument
0282			Person
0283	Playing Musical Instrument	Playing Musical Instrument	Person, Playing Musical Instrument
0284	Nighttime, Emergency Vehicle	Cityscape, Street	Cityscape, Road, Street, Nighttime
0285			
0286	Demonstration Or Protest		Explosion/Fire
0287			Person, Computer/TV
0288	Airplane flying, Boat Ship, Bridge	Airplane flying	Airplane, Airplane Flying, CU-Helicopters, Outdoor
0289		Chair	Person, Chair, CU-Talking
0290	Boat/Ship, Harbor	Boat/Ship	Boat/Ship, Waterscape
0291			Outdoor
0292	Female human face closeup	Female human face closeup	Female human face closeup

classifiers trained with different image and video features. This is followed by a post-processing stage utilizing the concepts' temporal and inter-concept co-occurrences. The results were rather promising and largely as expected. In particular, the SVM-based detectors obtained, expectedly, large accuracy improvements over the SOM-based baseline. We also performed a large set of experiments to be able to compare a variety of detector components in a systematic fashion. However, the detector-wise fluctuation in the results is so strong that we analyzed the components only using the averaged results which we consider to be more reliable.

In automatic search, the results again validate our earlier observations [13] that high-quality semantic concept detectors can be a considerable asset in automatic video retrieval. The mapping of concepts to queries is equally important, but how to do this optimally remains, however, an unresolved problem. In this work, the mapping of concepts to search queries was performed using the visual examples and a relatively naïve lexical matching approach. Using the visual examples works reasonably well, at least with the current concept ontology sizes, but a single-word-based lexical concept matching scheme is clearly insufficient.

ACKNOWLEDGMENTS

This work was supported by the funding of Academy of Finland for the *Finnish Centre of Excellence in Adaptive Informatics Research*. We thank the MediaMill and CU-VIREO teams for sharing their concept detectors with the TRECVID community.

REFERENCES

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

[2] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, Jorma Laaksonen, and Philip Prentis. PicSOM experiments in TRECVID 2007. In *Proceedings of the TRECVID 2007 Workshop*, Gaithersburg, MD, USA, November 2007. Available online at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[3] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, and Jorma Laaksonen. PicSOM experiments in TRECVID 2008. In *Proceedings of the TRECVID 2008 Workshop*, Gaithersburg, MD, USA, November 2008. Available online at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[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] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In *TREC Video Retrieval Evaluation Online Proceedings*. TRECVID, 2004.

[6] 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*, (in press), 2010.

[7] Marijn Huijbregts, Roeland Ordelman, and Franciska de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of the second international conference on*

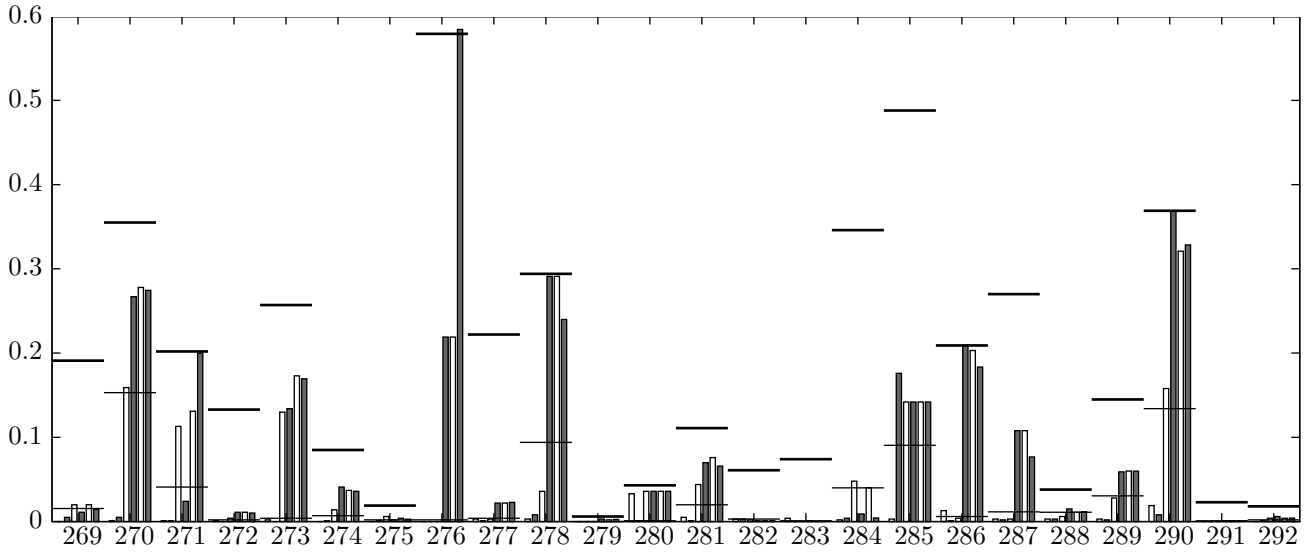

Fig. 4. The topic-wise MIAP results for our submitted automatic search runs 1, 2, 4, 6 and 8, in that order. In addition we show the additional run 8b as the last bar. The median and maximum values over all submissions are illustrated as horizontal lines.

Semantics And digital Media Technologies (SAMT), Lecture Notes in Computer Science, Berlin, December 2007. Springer Verlag.

[8] Stéphane Ayache and Georges Quénot. Video corpus annotation using active learning. In *Proceedings of 30th European Conference on Information Retrieval (ECIR'08)*, pages 187–198, Glasgow, UK, March-April 2008.

[9] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[10] D. Madigan A. Genkin, D. D. Lewis. BBR: Bayesian logistic regression software, 2005. Software available at http://www.stat.rutgers.edu/ madigan/BBR/.

[11] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, and Jorma Laaksonen. Exploiting temporal and inter-concept co-occurrence structure to detect high-level features in broadcast videos. In *Proceedings of 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008)*, pages 12–15, Klagenfurt, Austria, May 2008.

[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] Markus Koskela, Mats Sjöberg, and Jorma Laaksonen. Improving automatic video retrieval with semantic concept detection. In *Proceedings of 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, pages 480–489, Oslo, Norway, 2009. Springer Verlag.

[14] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. C. van Gemert, J. R. R. Uijlings, and et al. The mediamill TRECVID 2008 semantic video search engine. In *Proceedings of the TRECVID Workshop*, 2008.

[15] Yu-Gang Jiang, Akira Yanagawa, Shih-Fu Chang, and Chong-Wah Ngo. CU-VIREO374: Fusing columbia374 and VIREO374 for large scale semantic concept detection. In *Columbia University ADVENT Technical Report #223-2008-1*, August 2008.

[16] Kristina Toutanova and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 2000)*, pages 63–70, Hong Kong, October 2000.

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