PicSOM Experiments in TRECVID 2007

Markus Koskela, Mats Sjöberg, Ville Viitaniemi, Jorma Laaksonen
Adaptive Informatics Research Centre
Helsinki University of Technology, Finland

Philip Prentis
Faculty of Nuclear Sciences and Physical Engineering
Czech Technical University in Prague, Czech Republic

Abstract

Our experiments in TRECVID 2007 include participation in the high-level feature extraction, search, and video summarization tasks, using a common system framework based on multiple parallel Self-Organizing Maps (SOMs).

In the high-level feature extraction task, we applied a method of representing semantic concepts as class models on parallel SOMs, combined with external text search results. This year, we introduced a further post-processing stage in which the concepts' temporal and inter-concept co-occurrences were analyzed. We submitted the following six runs:

- `A_PicSOM_1_6`: Required visual baseline
- `A_PicSOM_2_5`: Visual features and text search
- `A_PicSOM_3_3`: Visual features using variable convolution and text search
- `A_PicSOM_4_4`: Visual features using variable convolution
- `A_PicSOM_5_2`: Visual features, text search, and temporal context based on training set
- `A_PicSOM_6_1`: Visual features, text search, and temporal context based on validation set

The results show that the temporal and inter-concept co-occurrence analysis improved the results considerably. On the other hand, inclusion of the text search worsened the results, leading to overall degradation of performance also on the subsequent runs. For this reason, we later executed additional runs in which the co-occurrence post-processing stage was employed without the text search.

In the search task, we submitted a total of six fully-automatic runs. In this year's experiments, we augmented the baseline ASR/MT search and content-based retrieval runs with high-level semantic concepts and pseudo relevance feedback. The overall settings for the six runs were as follows:

- `F_A_1_PicSOM_1_6`: Required text search baseline
- `F_A_1_PicSOM_2_5`: Required visual baseline
- `F_A_2_PicSOM_3_4`: Text search and visual features
- `F_A_2_PicSOM_4_3`: Text search and visual features, with pseudo relevance feedback
- `F_A_2_PicSOM_5_2`: Text search and visual features, semantic concepts
- `F_A_2_PicSOM_6_1`: Text search and visual features, semantic concepts, with pseudo relevance feedback

In this year's experiments, retrieval based on the visual features performed very poorly, and consequently the text baseline outperformed also the combined run with both visual and text features. In the further experiments, the inclusion of both the semantic concepts and pseudo relevance feedback resulted in performance improvement.

I. Introduction

In this paper, we describe our experiments for the TRECVID 2007 [1] evaluations. This year we participated in the high-level feature extraction, automatic search, and, for the first time, rushes video summarization. The basic system and methodology used in these experiments remains the same as in our previous participations in years 2005–2006. In the high-level feature extraction task, our main new addition to the system was a post-processing stage in which the temporal and inter-concept co-occurrences were analyzed. For our automatic search runs, we applied different combinations of external text search results, content-based retrieval based on visual (image and video) features, semantic concept modeling, and pseudo relevance feedback.

The rest of this notebook paper is organized as follows. The PicSOM system and the used visual and textual content descriptors are briefly described in Section II. Our experiments for the high-level feature extraction and fully automatic search tasks are described in Sections III and IV, respectively. The approach used in the video summarization task is briefly introduced in Section V and conclusions are presented in Section VI.

II. Indexing Video with PicSOM

The PicSOM system [2] 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) [3], each trained with separate feature data.

For video material, the index is based on a multimodal hierarchy for each individual video shot, which is considered as the main or parent object. The associated keyframes, the audio track, and ASR/MT text are linked as children of the parent object. Both the parent and children objects have one or more associated feature indices, and the relevance ratings and annotations are propagated within the object hierarchy. For a more detailed description of the used basic setup in our TRECVID evaluations, see [4], [5].

We extracted in total 15 video and 17 still image features from the Sound and Vision material. The still images were

keyframes extracted from the video shots in the master shot reference [6] using a heuristic algorithm. In this selection video frames are awarded for closeness to the temporal center of the shot, and penalized for distance from the calculated average image and for having big changes as compared to neighboring frames. The keyframe is selected as the frame with the highest score, the idea being that it should be close to the center, but at the same time be "typical" and not contain rapid movement, which could introduce e.g. motion blurring. No audio features were used in this year's experiments. The ASR/MT output was indexed separately using an external text search engine. All these features are briefly described in Sections II-A to II-C.

Separate 256×256-sized SOMs were trained for each of the video and image features. In this year's experiments, we actually trained two SOM indices for each feature: one using our original Tree-Structured SOM (TS-SOM) algorithm [2] and another using a triangular neighborhood kernel in the training of the TS-SOM.

A. Image features

For the keyframes, we extracted a large set of different features. First, six standard MPEG-7 descriptors, i.e. *Color Layout*, *Color Structure*, *Dominant Color*, *Scalable Color*, *Edge Histogram*, and *Region Shape*, were extracted using the MPEG-7 XM[1] reference software. For comparison, we also included our own implementations of four MPEG-7 descriptors, namely *Color Layout*, *Dominant Color*, *Scalable Color*, and *Edge Histogram*. All these descriptors were extracted globally from every keyframe.

In addition to the MPEG-7 features, we also extracted the following six non-standardized image features (see [5] for details): *Average Color*, *Color Moments*, *Texture Neighborhood*, *Edge Histogram*, *Edge Co-occurrence*, and *Edge Fourier*. These were calculated for five spatial zones of each image and the values concatenated to one image-wise vector.

Furthermore, we included a new image feature, *Interest Points*, based on histograms of interest point features. The interest points were detected using a combined Harris-Laplace and Difference-of-Gaussian detector, and SIFT features [7] were calculated for each interest point. Histograms of the SIFT features were then formed according to codebook vectors selected using the SOM as a clustering method. The size of the used SOM was 40×50 units, thus making the size of the codebook 2000 vectors.

Finally, we utilized a set of specific frame-level detectors for different purposes. For face detection, we used the detector included in Intel's OpenCV Library[2]. The detector is based on Haar-like features and a cascade of boosted tree classifiers. The face candidates returned by the OpenCV detector were pruned by using a simple skin color detector in the YCbCr color space [8]. This face detector was utilized in all three tasks. Additionally, we used a greyscale shot detector based on the *Average Color* feature in the search task, as well as a "junk" shot detector (for color bars and empty frames) and a motion detector in the rushes summarization task [9].

B. Video features

On the video shot level, we used the MPEG-7 *Motion Activity* descriptor (MA) and temporal versions of both our own implementations of the four MPEG-7 descriptors and the six non-standard still image features described above. The temporal versions of the still-image features were calculated by averaging over the frames contained within five non-overlapping temporal video slices in the shot. Each frame was also divided into five spatial zones. This way we obtained feature vectors that describe changes in the still-image features over time in different spatial areas of the video.

C. Text features

Both the automatic speech recognition (ASR) output [10] and the machine-translated (MT) English version of it were separately indexed and queried using the Apache Lucene[3] text search engine. The ASR output (in Dutch) was used in the high-level feature extraction task. Each shot was indexed separately, i.e. the ASR outputs were not spread within the shots' temporal neighborhoods. In the search task, only the machine-translated English index was queried using the provided English language topic descriptions. In this task, the text search was a combination of shot-wise indices and using a three-shot-radius neighborhood to temporally spread the ASR/MT outputs. The Snowball stemmers[4] were used for both Dutch and English (Porter2), with the included stop word lists.

III. High-level Feature Extraction

For the high-level feature extraction task this year, we incorporated a temporal and inter-concept co-occurrence analysis step to our existing SOM-based method for concept modeling. The basic method is based on modeling probability densities of the concepts using kernel-based estimation of discrete class densities over the SOM grids. See [4], [11] for more details. All 36 high-level concepts are detected using the same procedure based on the concept-wise ground-truth annotations gathered by the organized collaborative annotation effort [12].

Table I gives an overview of the high-level feature extraction runs. The columns refer to the inclusion of text features, the size of the used kernels (common width or optimized for each concept separately), and the two methods used for temporal and inter-concept co-occurrence analysis.

The first run is the required baseline where only visual features are used. Run 2 combines the visual features with the external text search results in Dutch. The Apache Lucene search engine was used for the text-based search, with the concept-wise query terms consisting of the most common terms in the ASR output of the the positive training examples in the development set. In these runs, the radius of the

[1] http://www.lis.ei.tum.de/research/bv/topics/mmdb/e_mpeg7.html
[2] http://www.intel.com/technology/computing/opencv/
[3] http://lucene.apache.org/
[4] http://snowball.tartarus.org/

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

#	run id	text feat.	kernel size common	kernel size variable	temporal train	temporal val	mean InfAP
1	A_PicSOM_1_6		•				0.0621
2	A_PicSOM_2_5	•	•				0.0550
3	A_PicSOM_3_3	•		•			0.0526
4	A_PicSOM_4_4			•			0.0612
5	A_PicSOM_5_2	•	•		•		0.0747
6	A_PicSOM_6_1	•	•			•	0.0772
7	additional run		•		•		0.0792
8	additional run		•			•	0.0804
9	additional run		oracle				0.0636
10	additional run			oracle			0.0694
11	additional run		•		oracle		0.0926

triangular-shaped kernel in the class density estimation over the SOM surfaces is always 8 units. This value is the result of two-fold cross-validation on the development data.

For runs 3 and 4, the kernel size is optimized for each concept separately, again with two-fold cross-validation on the development set. This resulted in kernel radii varying from 5 to 30 units. Otherwise, the runs differ only on the inclusion of the text features.

The runs 5 and 6 include a post-processing step based on the temporal co-occurrence of the concepts. The used technique is described in Section III-B and in more detail in [13]. Run 2 was selected as the starting point for the temporal post-processing for the submitted runs 5 and 6. However, as the text features turned out to degrade the results compared to the visual baseline (run 1), the same experiments were later reproduced with the baseline run 1 as the starting point. These results are shown in Table I as additional runs 7 and 8.

In order to complement our submitted experiments with varying kernel sizes, we also carried out additional experiments in which the radius of the kernel was set to all values between 5 and 30 map units. Based on these experiments, we composed two additional runs, numbered 9 and 10, in which the common and concept-wise kernel sizes, respectively, are given the optimal or oracle values. For the common kernel radius parameter (run 9), the oracle value was either 10 or 11 map units.

Finally, in run 11 we have the same post-processing based on temporal and inter-concept co-occurrence as in runs 5–8, but using optimal (oracle) selection of the post-processors.

A. Feature selection and weighting

As in our previous TRECVID experiments, the set of features (and associated SOM indices) was selected for each concept separately using a greedy feature selection scheme [4]. In these experiments, the pool of potential feature indices was, however, doubled as both a standard and a triangular-neighborhood SOM were trained for each feature. Both SOMs were included separately in the pool, so it was possible that both indices would be selected for the same concept.

The feature selection algorithm resulted in 9.4 feature indices per concept on average. The most frequently selected features were *Interest Points*, the MPEG-7 XM version of the still-image *Edge Histogram*, and the temporal versions of *Edge Fourier* and *Texture Neighborhood*.

The text search results were included as external features into the feature fusion stage of the PicSOM system. However, as the inherent feature weighting mechanism of the system is not able to automatically weight externally provided features, the corresponding concept-wise weights were optimized using the development set. The same optimization was performed also for the OpenCV face detection results, but this resulted in non-zero weights only for concepts *face* and *person*, neither of which were included in the set of concepts evaluated at NIST.

B. Temporal and inter-concept co-occurrence

For runs 5 and 6 the detection results of run 2 were post-processed to take advantage of both temporal and inter-concept co-occurrences. N-gram models of order 1 to 4 were used as the temporal component. The inter-concept co-occurrences were put to use by clustering the 36-dimensional concept detection vectors into 16 clusters with the LBG-algorithm and modeling each cluster separately. In addition to clustering the detection vectors of single time instants, we also generated clusterings based on average detection vectors in temporal windows of various lengths.

We generated 15 separate post-processors applying the outlined techniques in various combinations and with various model parameters (including a do-nothing baseline). Each post-processor was trained using six-fold cross-validated detections of the development set. For the two runs we tried two methods of selecting a post-processor for each concept. For run 5 we chose the post-processor with maximum performance in the training set. For run 6 we performed a separate validation experiment of the post-processors by training with half of the development set and validating in the other half.

C. Results

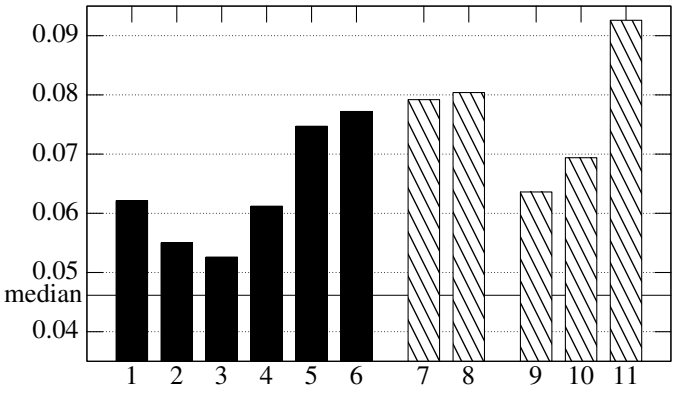

Fig. 1. Mean InfAP values for our runs in the high-level feature extraction task; the submitted runs shown as black bars. The median of all submitted runs in also shown.

Figure 1 illustrates the mean inferred average precision (InfAP) [14] values of our runs in the high-level feature extraction task. The highest mean InfAP score of our submitted

runs was 0.077 obtained with run 6. The median and maximum over all 105 submissions were 0.046 and 0.132, respectively.

First of all, from the runs 1–4 it can be observed that the ASR text features degrades the overall results, both when using common and variable kernel sizes. This result is contrary to our results in last year's high-level feature extraction experiments [5], highlighting the significance of the difference between the types of used video materials. Furthermore, the concept-wise optimization of the kernel radius also degraded the overall results, making the visual baseline (run 1) the highest scoring run among the runs 1–4.

We analyzed the effect of the kernel radius parameter further with two additional (non-submitted) runs. In run 9, the common value for the kernel radius is set to its optimal or oracle value of 10 units. The corresponding mean InfAP value, however, shows only a modest improvement over the visual baseline. In run 10, the concept-wise kernel radii are set to their corresponding oracle values in the range of 5 to 30 map units. Here, it can be seen from the mean InfAP value that this results in a notable improvement compared to both runs 1 and 4. The two-fold cross-validation performed on the development set was not, however, able to select these optimal values, or even improve over the common kernel size shared by all the concepts.

The temporal and inter-concept co-occurrence post-processing shows rather consistent improvement in mean InfAP. Runs 5 and 6 show 36% and 40% increase over their starting point (run 2). These results were, however, burdened by the performance degradation caused by the inclusion of the text-based search results. Therefore, we reproduced the analysis using the visual baseline (run 1) as the starting point. These additional runs 7 and 8 show similar behavior, resulting in increases of 27% and 29%, respectively, in mean InfAP. Furthermore, in run 11 the result is shown if we select the post-processors optimally for each concept. This produces an increase of 48% over the visual baseline.

The concept-wise results are illustrated in Figure 2, which shows the InfAP values of our submitted runs for all concepts. The best one among our submitted runs was substantially over the median with concepts *weather* (3), *office* (5), *meeting* (6), *waterscape/waterfront* (17), *animal* (26), *computer/tv screen* (27), *car* (30), *boat/ship* (33), and *people marching* (35). For the concept *meeting*, our run 6 resulted in the highest InfAP over all runs submitted in the TRECVID 2007 evaluations.

IV. Automatic Search

For the search task, we submitted six automatic runs summarized in Table II. All runs were trained only on common TRECVID development data, thus qualifying as type A runs. The retrieval technique is similar to the one used in our previous TRECVID submissions. The general idea is to combine retrieval based on SOM indices trained with visual features with external text-based search and semantic concept models. The runs numbered 1 and 2 constitute the required baseline runs using only text-based search and visual features, respectively. Runs 3 and 4 combine the text and visual modalities, with pseudo relevance feedback (PRF) applied in run 4. The semantic concept models are then introduced in runs 5 and 6, again with PRF in the latter run.

TABLE II
An overview of our submitted search task runs.

#	run id	text	visual	concepts	PRF	MAP
1	F_A_1_PicSOM_1_6	•				0.0122
2	F_A_1_PicSOM_2_5		•			0.0014
3	F_A_2_PicSOM_3_4	•	•			0.0085
4	F_A_2_PicSOM_4_3	•	•		•	0.0115
5	F_A_2_PicSOM_5_2	•	•	•		0.0191
6	F_A_2_PicSOM_6_1	•	•	•	•	0.0220

A set of 8 visual (4 image and 4 video) features was selected as a common feature set for all topics based on their performance in the feature selection process of the high-level feature extraction task (see Section III-A). For analyzing the topic-wise text descriptions, the Stanford part-of-speech tagger[5] [15] was used. The nouns and verbs of each textual description were used as the text search queries for the Apache Lucene text search engine. The relative weight of the text-based search results were increased for topics containing proper nouns in the textual description. This applied only to topic 219.

A. Semantic concept matching

The search topics were matched with suitable semantic concepts to facilitate concept model based retrieval in runs 5 and 6. In these experiments, we limited the selection of available concepts to the 36 high-level features detected in the high-level feature extraction task and additional keyframe-based detectors for faces and greyscale shots (see Section II-A). The semantic concepts were modeled on the same 8 common features used for content-based retrieval. The matching of semantic concepts to queries was based on a lexical analysis of the topic-wise textual descriptions.

B. Pseudo relevance feedback

This year we included an optional pseudo relevance feedback (PRF) stage to re-rank the retrieval results in the automatic search task. For each topic, one round of PRF was carried out, with the 20 best-scoring shots marked as additional positive examples. This processing step was applied in runs 4 and 6.

C. Results

The MAP scores for all our search runs are listed in Table II and the topic-wise results are shown in Figure 3. The runs 1 and 2 are of the required baseline type, among which the maximum and median over all submissions were 0.043 and 0.004, respectively. Of our baselines, the ASR/MT text search run performed considerably better than the run using only visual features.

All our runs used only the common TRECVID development data, thus qualifying them as type A runs. Over all type A automatic search submissions, the maximum and median values

[5]http://nlp.stanford.edu/software/tagger.shtml

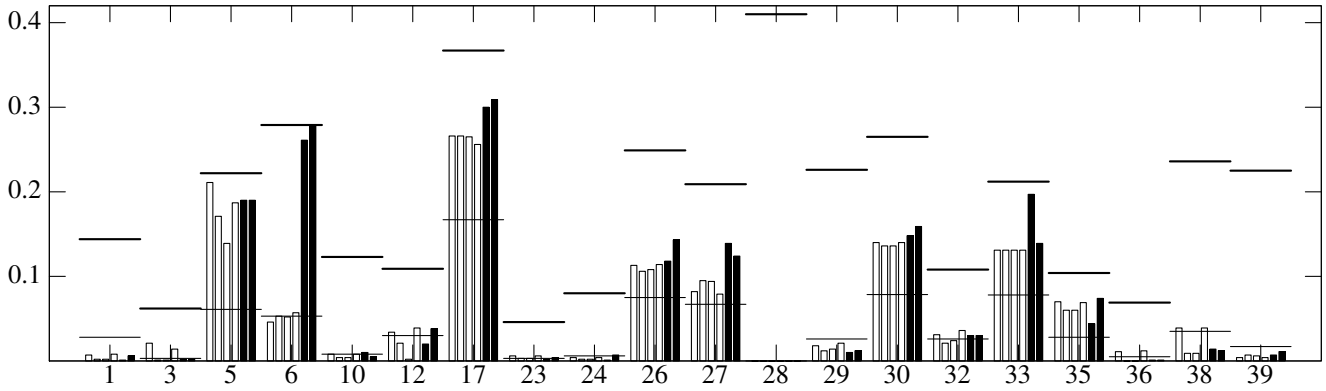

Fig. 2. The concept-wise InfAP results of our submitted runs for each evaluated concept. The runs including the temporal co-occurrence analysis are shown as black bars. The median and maximum values over all submissions are illustrated as horizontal lines.

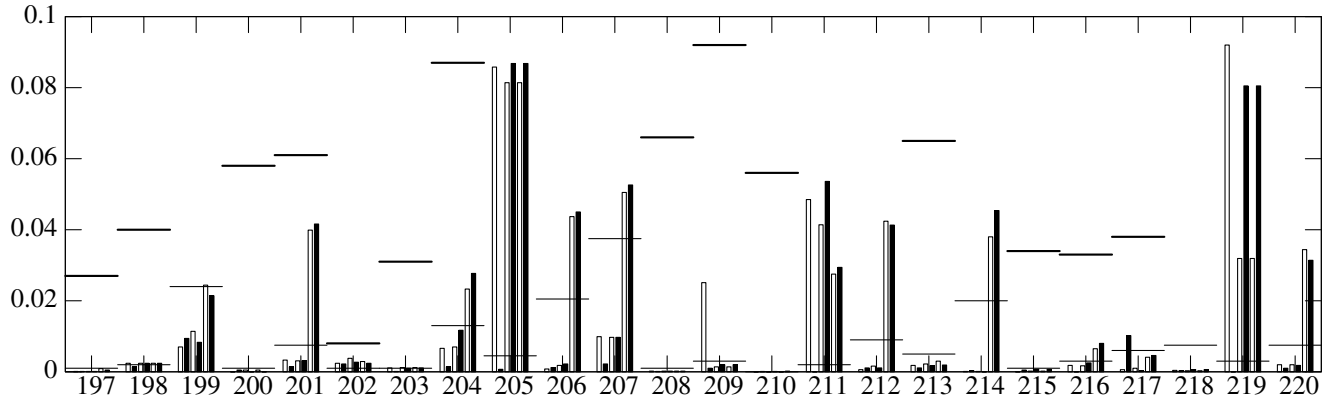

Fig. 3. The topic-wise AP results for our submitted automatic search runs, with runs 2, 4 and 6 drawn as black bars. The median and maximum values over all submissions are illustrated as horizontal lines; not all maximums are visible in the figure.

were 0.087 and 0.014, respectively. Run 3 is a combination of the text search and content-based retrieval. Due to the poor performance of the latter modality, the resulting MAP does not reach the text baseline. Next, the inclusion of the semantic concept models in run 5 improves the results. From the topic-wise results it can be observed that the concept models help significantly with about one third of the topics. Small improvements can also be observed by using pseudo relevance feedback, both with and without the semantic concepts.

V. Video Summarization

We also participated in the BBC rushes summarization task [16] using an approach consisting of initial shot boundary detection followed by shot similarity assessment and pruning, with both stages implemented using multiple SOMs. The approach is described in more detail in [9]. First, we apply our shot boundary detection algorithm [17], [5] to the rushes videos. For each video, this provides us with lists of shots, which are used in the following stages as basic units of processing. We detect and remove unwanted "junk" shots (color bar test screens, empty frames) from the videos, and apply face detection and motion activity estimation. Next, we compute the visual similarities between all pairs of shots and remove overly similar shots. Each remaining shot is then represented in the summary with a separately selected one-second clip, with the audio track not included. The selected clips are combined using temporal ordering and fade-outs and fade-ins from black.

VI. Conclusions

This was our third year participating in the TRECVID evaluations. The basic functionality of the system has been the same in all experiments, and we have introduced new additions to the system each year.

Extracting high-level features using the SOM-based approach is efficient and highly scalable to large ontologies due to the modeling of the concept densities in low-dimensional spaces using non-parametric kernel-based density estimation. It shows relatively good performance, although not quite reaching the level of computationally more complex discriminative methods such as SVMs. The method is not particularly sensitive to width or form of the kernel function. The common size is a stable parameter, but it is possible to improve the results using concept-wise values, at least in the optimal case.

The high-level feature extraction results can be further improved with different techniques, such as using auxiliary

concepts (in last year's experiments [5]) and analyzing temporal and inter-concept co-occurrences. This analysis technique introduced here is still rather preliminary and undoubtedly has room for improvement. Despite this it proved to be very promising giving a significant increase in retrieval performance.

Due to the change in the type of video material analyzed, there were no useful annotations for the full LSCOM [18] ontology. This is because we discovered in initial experiments that by using the previous year's ground-truth directly, the resulting concept models were not satisfactory. As a result, we did not employ any auxiliary concepts in the high-level feature extraction task and mapped only the set of 36 high-level features into the topics in the search task. Despite the smaller set of concepts available, the semantic concept models were again found to be useful in automatic search,

We also included a pseudo relevance feedback option to the system this year. The results indicate that it generally leads to small improvements in search performance. After the submissions, we tested including pseudo relevance feedback to the high-level feature extraction runs as well, with very similar results.

ACKNOWLEDGMENTS

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*.

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] 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.

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

[4] 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.

[5] Mats Sjöberg, Hannes Muurinen, Jorma Laaksonen, and Markus Koskela. PicSOM experiments in TRECVID 2006. In *Proceedings of the TRECVID 2006 Workshop*, Gaithersburg, MD, USA, November 2006. Available online at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

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

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

[8] D. Chai and K. N. Ngan. Face segmentation using skin-color map in videophone applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(4):551–564, June 1999.

[9] Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Ville Viitaniemi, and Hannes Muurinen. Rushes summarization with self-organizing maps. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 45–49, Augsburg, Germany, September 2007. ACM Press.

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

[11] Markus Koskela and Jorma Laaksonen. Semantic concept detection from news videos with self-organizing maps. In Ilias Maglogiannis, Kostas Karpouzis, and Max Bramer, editors, *Proceedings of 3rd IFIP Conference on Artificial Intelligence Applications and Innovations*, pages 591–599, Athens, Greece, June 2006. IFIP, Springer.

[12] Stéphane Ayache and Georges Quénot. TRECVID 2007 collaborative annotation using active learning. In *Proceedings of the TRECVID 2007 Workshop*, Gaithersburg, MD, USA, November 2007.

[13] 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)*, Klagenfurt, Austria, May 2008. Accepted.

[14] 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.

[15] 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.

[16] Paul Over, Alan F. Smeaton, and Philip Kelly. The TRECVID 2007 BBC rushes summarization evaluation pilot. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 1–15, New York, NY, September 2007. ACM Press.

[17] 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)*, pages 770–779, Aalborg, Denmark, June 2007.

[18] 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.