Exploiting Temporal and Inter-Concept Co-Occurrence Structure to Detect High-Level Features in Broadcast Videos

Ville Viitaniemi, Mats Sjöberg, Markus Koskela, Jorma Laaksonen
Adaptive Informatics Research Centre
Helsinki University of Technology, Finland
{ville.viitaniemi,mats.sjoberg,markus.koskela,jorma.laaksonen}@tkk.fi

Abstract

In this paper the problem of detecting high-level features from video shots is studied. In particular, we explore the possibility of taking advantage of temporal and inter-concept co-occurrence patterns that the high-level features of a video sequence exhibit. Here we present two straightforward techniques for the task: N-gram models and clustering of temporal neighbourhoods. We demonstrate the usefulness of these techniques on data sets of the TRECVID high-level feature detection tasks of the years 2005-2007.

1. Introduction

Extracting semantic concepts from multimedia data has attracted a lot of research attention recently [3]. The main aim is to facilitate semantic indexing and concept-based retrieval of multimedia content. The leading principle is to build semantic representations by extracting intermediate semantic levels from low-level features. Recently, the introduction of large-scale multimedia ontologies (e.g. [2]) as well as large manually annotated datasets have enabled generic analysis of multimedia content as well as an increase in multimedia lexicon sizes by orders of magnitude.

One of the tasks in the annual TRECVID video retrieval evaluation [6] is to detect the presence of predefined high-level features (HLF)—such as "sports", "meeting" or "urban"—in broadcast videos that are already partitioned into shots. The predominant approach to detecting HLFs is to treat the problem as a generic supervised learning problem. This automatic approach is scalable to large numbers of features. The training data is used to learn independent models of different concepts over low-level feature distributions.

It is almost self-evident that the HLFs in videos have temporal structure, for example the HLFs—also called concepts—of subsequent shots are likely to be similar. The alternation of concepts might also exhibit some characteristic temporal patterns. The existence of such temporal structure is evidenced in the mutual information between the detections of the same concept at close-by time instants, evaluated in [7] for a time lag of one shot and by ourselves also for longer time spans (not described here).

However, it is not so self-evident that the concept detection accuracy can be improved by using temporal characteristics of the video streams. There might be several hurdles. When detecting features of novel video material, only detector outputs of probabilistic nature are available, not binary oracles indicating the actual presence of the features. Often the detections of some features are very inaccurate. In their earlier study Yang and Hauptmann [7] arrived at the result that temporal models conditional to oracle detections significantly improved the detection result but with real detections the temporal model brought very little improvement. They pointed out the possibility that temporally close shots might usually be similar by their low-level features, and the corresponding detector outputs thus very correlated. The resulting "miss-one-miss-all" phenomenon would make temporal smoothing less effective.

Another nearly obvious characteristic of the high-level features is that the features exhibited by a video shot are mutually dependent. For instance, the feature "snow" almost always implies "outdoor", whereas concepts like "sports" and "weather forecast" are practically mutually exclusive. It has been experimentally found very beneficial to exploit concept co-occurrence for HLF detection [4, 5].

Despite the potential obstacles, it is reasonable to believe that with an adequate amount of training material the estimation of an accurate generative model behind the observations should lead to optimal results. However, the training material used in TRECVID campaigns in years 2005-2007 is limited when it comes to modelling the temporal and co-occurrence structure of the videos. Even though the total number of shots is rather large, within a given video the temporal and co-occurrence structures are likely to be similar. Thus the number of independent training samples

978-0-7695-3130-4/08 $25.00 © 2008 IEEE
DOI 10.1109/WIAMIS.2008.50

for the temporal and co-occurrence models would be much smaller, perhaps of the order of the number of videos in the training material. This in turn is of order 100 in the TRECVID data, certainly not a large number of training samples to train a complex model.

In this paper we propose a set of simple post-processing techniques to model the temporal and inter-concept co-occurrences as an add-on, when the concepts have already been detected without considering these issues. We combine the N-gram technique for intra-concept temporal modelling, and a simple clustering technique that takes advantage of inter-concept temporal and instantaneous co-occurrences. These techniques can be seen as a superset of the intra-concept bigram technique of [7]. We demonstrate the usefulness of the proposed techniques within two data sets: the TRECVID 2005/2006 HLF task development videos and all of TRECVID 2007 HLF task videos.

2. Techniques for exploiting temporal and inter-concept co-occurrences

In this section we describe techniques that take into account the temporal and inter-concept co-occurrence as a post-processing step. The techniques operate on a stream of K-tuples corresponding the concept detector outputs for the sequential video shots, where K is the number of the concepts detected. The presented methods thus ignore the absolute timing and duration of the video shots, preserving just the ordering of the shots. Methodologically, we propose two types of approaches: N-gram models and clusterings of temporal neighbourhoods.

2.1. N-gram models

The N-gram model was applied to each concept individually. In the following, $c_n \in \{0, 1\}$ is indicator of the occurrence of the concept to be detected at time instant n, and $s_n \in \mathbf{R}$ is the output of the corresponding concept detector. $H_n(N)$ denotes the recursive prediction history known at time instant n, extending $N-1$ steps backwards in time:

$$H_n(N) = \{\hat{p}(c_{n-i}|s_{n-i}, H_{n-i}(N))\}_{i=1}^{N-1} \quad (1)$$

Using this notation, we can write the recursive N-gram model as

$$\hat{p}\left(c_n|s_n, H_n(N)\right) \propto \hat{p}(s_n|c_n)\hat{p}\left(c_n|H_n(N)\right). \quad (2)$$

In this recursive model

$$\hat{p}\left(c_n|H_n(N)\right) = \sum_{c_{n-1}} \cdots \sum_{c_{n-N+1}} \quad (3)$$

$$p_0(c_n|c_{n-1}\cdots c_{n-N+1}) \prod_{i=1}^{N-1} \hat{p}\left(c_{n-i}|s_{n-i}, H_{n-i}(N)\right)$$

Here p_0 is the marginalised N-gram probability estimated from the training data. The N-gram model was initialised in the beginning of each video by using models of lower order, e.g. bigram model at the second time instant. The conditional distributions of detector outputs $\hat{p}(s_n|c_n)$ were modelled as exponential distributions whose parameters were estimated from the training data by means of maximum likelihood.

In addition to this causal model, we also formed the corresponding anticausal model that is obtained by reversing the time flow. The causal and anticausal models were combined by logarithmic averaging of the model outcomes.

2.2. Clusterings of temporal neighbourhoods

The N-gram model was augmented with information C_n that was obtained by clustering the baseline detector outputs within temporal neighbourhoods around the prediction time instant n. The clustering was based simultaneously on all the K concepts, i.e. 36 or 39 in the experiments reported here. The LBG algorithm with 16 clusters was used. The cluster information was combined with the N-gram model by estimating the N-gram model separately for each cluster, resulting in models for for $p_0(c_n|C_n, c_{n-1}\cdots c_{n-N+1})$. The cluster-specific detector outcome distribution $\hat{p}(s_n|C_n, c_n)$ was modelled as a linear interpolation between the global logistic model and a logistic model estimated for each cluster separately.

Several different clusterings were combined by taking logarithmic averages of the detection probability estimates based on each clustering. The different clusterings resulted from neighbourhoods of different time spans.

3. Experiments

A set of experiments were performed with two data sets: the development videos of the high-level feature (HLF) detection task of TRECVID 2005/2006, and both the development and test data of the TRECVID 2007 HLF task. The results on these two data sets seem to point to the same direction: there is some advantage of using the methods outlined in Section 2. However, for different data sets somewhat different techniques turned out to be the most beneficial.

For both sets of experiments, we post-processed the detector outputs of our PicSOM video analysis system that were used when participating in the TRECVID HLF detection tasks [5, 1]. For this post-processing we employed the N-gram and clustering techniques of Section 2 in various combinations and with varying model parameters. The post-processors as well as the original detectors operate in a supervised manner, i.e. both data sets were partitioned into test and training subsets. The estimates of the detector

outputs for the training set required by the post-processing techniques were obtained by using cross-validation.

Each post-processed detection stream was evaluated for the individual concepts using the same metric that was used in the TRECVID evaluation of the corresponding year: (non-interpolated) average precision (AP) for the 2005/2006 data, and inferred average precision (infAP) [8] for the 2007 data. For the 2005/2006 development set all the 39 detected concepts were evaluated. For the 2007 data we detected 36 concepts, but only 20 were evaluated for the test data by the TRECVID organisers. In the following, we report some highlights of the detection results. Many interesting properties appear in the concept-wise detection results, but due to the space limitations here we mainly concentrate on the performances averaged over all of the evaluated concepts.

3.1. The 2005/2006 development data

The first set of experiments was performed on TRECVID 2005/2006 HLF development data (the same data was used in both years). The data consisted of 137 videos segmented into approximately 44 000 video shots, consisting mostly of news broadcasts in English, Arabic, and Chinese. The data was split approximately evenly into training and test halves. The detector outputs for the training half of the data were estimated by 7-fold cross validation. This data was completely annotated with 39 concepts.

Table 3.1 exemplifies some concept-wise detection results in the test set. Each row of the table contains the baseline detection result for that concept and lists the improvement percentages for the various alternative post-processors (not all post-processors are tabulated). The methods "2-gram" and "4-gram" use solely the temporal structure within detections of individual concepts, "co" utilises clustering based on the instantaneous co-occurrence of concepts, whereas in "cl" the clustering is based on a larger temporal neighbourhood. Next, "cl+co" combines the previous two techniques and "cl+co+2-gram" adds a bigram model to this. The last two columns show the performances of the methods that were best in the training and test sets ("oracle"), respectively. On the last row of the table we report the average performance (MAP) over all the 39 concepts.

In many cases, the proposed post-processing techniques are beneficial. Often the combination of techniques is more useful than one single technique. The relative ordering of the techniques varies from concept to concept. Within individual concepts the improvement percentages show larger variation than in the average performance. Both of these two characteristics are partly explained by the statistical fluctuations in a small sample. However, partially this testifies of the genuine differences of concepts as the performances in training and test sets are rather well correlated. This facilitates method selection based on training set performance. However, this correlation is not perfect, as evidenced by the difference between the last two columns of the table.

3.2. TRECVID 2007 experiments

In the 2007 TRECVID experiments [1] a different video data set consisting of a wide variety of Dutch television programming, such as news magazines, documentaries and educational shows was used. The full data set contains 219 videos segmented into approximately 36 000 video shots. In this experiment we generated 15 separate post-processors. Each post-processor was trained using 6-fold cross-validation on the training set. For each concept we tried two different methods of selecting the best post-processor: *method 1*: selecting the one with maximum performance in the training set, or *method 2*: by a separate validation experiment training with one half of the original training set and validating in the other.

In Fig. 1, the mean infAP results over all 20 evaluated concepts are summarised. The first four bars (B1 through O1) show the results of using visual features as the baseline, the last four bars (B2 through O2) use both visual and textual features. Our textual features did not work well with the 2007 data. The first bar in both groups (B1, B2) show the baseline performance of our PicSOM system, without taking advantage of temporal and inter-concept co-occurrences. Bars T1 and T2 show the results when using post-processors selected with *method 1* in addition to the baseline PicSOM features. Bars V1 and V2 correspond to validation by splitting the training set, i.e. *method 2*. Finally bars O1 and O2 show the performance that would be achieved with an optimal ("oracle") selection of post-processors. The median of all submitted runs from all groups is also shown for comparison.

From Fig. 1 it is quite evident that the temporal and inter-concept co-occurrence methods strongly improve upon the baseline results of PicSOM. Of the two post-processor selection methods, the method employing a validation set seemed to perform slightly better.

4. Conclusions

In this paper we reported a set of experiments that clearly indicate that using both temporal and inter-concept co-occurrences can be beneficial for detection of concepts from news videos. Temporal information can be found both in the occurrence history of a single concept and the occurrence history of other concepts. These temporal information sources and the inter-concept co-occurrence information seem to be partially complementary for many of the

Table 1. Examples of concept detection results using the TRECVID 2005/2006 data.

concept	baseline AP	2-gram	4-gram	co	cl	cl+co	cl+co+2-gram	best in train	oracle
face	0.790	-0.6%	-1.4%	+12.1%	+5.8%	+11.1%	+9.0%	+12.1%	+12.1%
maps	0.241	+1.7%	+1.5%	+39.8%	+0.9%	+35.8%	+32.0%	+32.0%	+39.8%
people-marching	0.024	+38.7%	+81.5%	+57.1%	+49.4%	+64.7%	+208%	+208%	+208%
court	0.067	+1.8%	+1.9%	-47.1%	+27.0%	-11.5%	-21.6%	-21.6%	+27.0%
outdoor	0.385	-4.4%	-6.5%	+12.7%	+20.0%	+22.3%	+21.2%	+22.3%	+22.3%
crowd	0.151	-0.7%	-2.9%	+30.7%	+28.9%	+36.0%	+50.3%	+36.0%	+55.4%
military	0.067	+69.0%	+96.5%	+10.0%	+47.1%	+41.1%	+123%	+123%	+123%
average	0.176	0.184	0.185	0.185	0.188	0.191	0.202	0.203	0.211

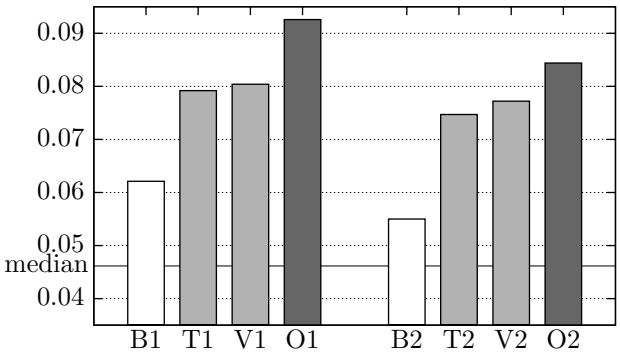

Figure 1. Mean InfAP values for the experiments on the TRECVID 2007 data.

detected concepts. For simplicity, we regarded here the original detections as an ordered symbol stream. Taking the detailed temporal structure—e.g. shot durations—into account could provide still more of useful information.

The presented experiments point to a different direction than the experiments by Yang and Hauptmann [7] with the TRECVID 2005/2006 development data. They did not find the temporal smoothing to be useful in their experiments with TRECVID data. However, our results are in line with the improvement reported in [4].

The usefulness of the outlined individual technical alternatives varies strongly between concepts. Some techniques are even harmful for some concepts. The two data sets of the experiments were somewhat different in this respect. We presented two methods of selecting between the techniques. Even if these methods seem to work to some extent, the results are not fully satisfactory, as evidenced by comparison with oracle selections of techniques. It is expected that the results could be improved by using more rigorous cross-validation.

The employed techniques are rather heuristically chosen and rudimentary. It seems likely that the information could be better exploited by using more rigorous and principled techniques. For example, temporal techniques routinely applied in speech recognition could be worth investigating also in this context. However, one has to keep in mind that the small number of temporal co-occurrence training samples in our experiments (and in the TRECVID tasks) seriously limits the complexity of models that can be estimated from the data.

It would be rather straightforward to refine also the techniques presented here. One obvious direction to look at is selection of variables on which the clusterings are performed. This could be done separately for each target concept based on the observed dependencies between the concept detectors. Currently the clusterings are dominated by reliably detected concepts that do not necessarily give much information about the target concepts.

References

[1] M. Koskela, M. Sjöberg, V. Viitaniemi, J. Laaksonen, and P. Prentis. PicSOM experiments in TRECVID 2007. In *Proc. of the TRECVID 2007 Workshop*, Gaithersburg, MD, USA, Nov. 2007.

[2] M. Naphade, J. R. Smith, J. Tešić, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[3] M. R. Naphade and T. S. Huang. Extracting semantics from audiovisual content: The final frontier in multimedia retrieval. *IEEE Transactions on Neural Networks*, 13(4):793–810, July 2002.

[4] S. Petrov, A. Faria, P. Michaillat, A. Stolcke, D. Klein, and J. Malik. Detecting categories in news video using acoustic, speech and image features. In *TRECVID Online Proceedings*. TRECVID, Nov. 2006.

[5] M. Sjöberg, H. Muurinen, J. Laaksonen, and M. Koskela. PicSOM experiments in TRECVID 2006. In *Proc. of the TRECVID 2006 Workshop*, Gaithersburg, MD, USA, Nov. 2006.

[6] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proc. of ACM MIR '06*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[7] J. Yang and A. Hauptmann. Exploring temporal consistency for video retrieval and analysis. In *Proc. of ACM SIGMM Int. Workshop on MIR*, Santa Barbara, CA, Oct. 2006.

[8] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proc. of CIKM2006*, Arlington, VA, USA, Nov. 2006.