Semantic Concept Detection from News Videos with Self-Organizing Maps*

Markus Koskela[1] and Jorma Laaksonen[2]

[1] Centre for Digital Video Processing, Dublin City University, Ireland
markus.koskela@computing.dcu.ie
[2] Adaptive Informatics Research Centre, Helsinki University of Technology, Finland
jorma.laaksonen@hut.fi

Abstract. In this paper, we consider the automatic identification of video shots that are relevant to a given semantic concept from large video databases. We apply a method of representing semantic concepts as class models on a set of parallel Self-Organizing Maps trained with multimodal low-level features. The presented experiments were conducted using a set of 170 hours of video containing recorded television news programs.

1 Introduction

Matching semantic concepts and visual data has attracted a lot of research attention recently in order to facilitate semantic indexing and concept-based retrieval of multimedia content. Traditional example-based retrieval via relevance feedback or other methods can be enriched with semantic concept models that have been trained off-line with considerably more positive and negative examples than what are available on-line for an ordinary image or video query. For producing large-scale semantic concept models of visual data, the predominant approach is to treat the problem as a generic learning problem in which existing sets of training data is used to learn models of different concepts over low-level feature distributions. This is due to scalability requirements, as a comprehensive visual lexicon needs models for hundreds or thousands of concepts.

In this paper, we study the problem of general semantic concept detection from news videos by utilizing a hierarchical approach to indexing video and by extracting multiple parallel features from the different data modalities. A set of Self-Organizing Maps (SOMs) is then trained on these features to provide a common indexing structure across the different modalities. The rest of the paper is organized as follows. The use of SOMs for indexing video and the used multimodal features are briefly described in Section 2. In Section 3 we discuss the use of parallel low-level SOM indices in modeling semantic concepts. A set of experiments in high-level concept detection on the TRECVID 2005 news video data are described in Section 4, and conclusions are presented in Section 5.

* 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*.

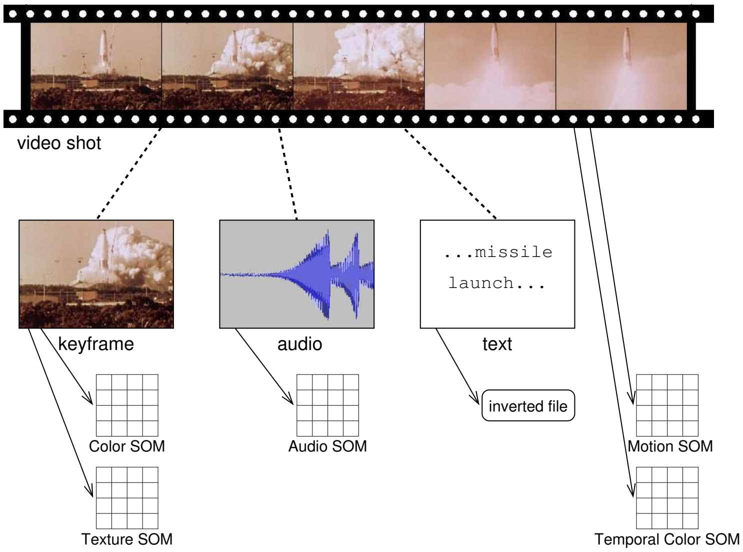

Fig. 1. A hierarchical view on video data and associated multimodal feature indices.

2 Indexing Video Shots with Self-Organizing Maps

The Self-Organizing Map (SOM) [1] is a powerful tool for exploring huge amounts of high-dimensional data. It defines an elastic, topology-preserving grid of points that is fitted to the input space. In a typical data mining, visualization, or information retrieval application, a SOM is trained in a fully unsupervised mode, using a large batch of training data. Yet, it is often known that the data contain some semantically related object groupings or classes, and there are available subsets of vectors belonging to such user-defined classes. Such a set of vectors can be mapped on a trained SOM by finding the best matching unit (BMU) for each vector in the set. These "hits" over the units of the SOM surface form a discrete probability distribution which characterizes the object class. Different distributions can be obtained by using different feature extraction techniques, leading to different representations of the same data items.

2.1 Indexing Hierarchical Objects

The PicSOM system [2] is a general framework for research on content-based indexing and retrieval of visual objects. An extension to PicSOM for indexing any multi-part and multimodal objects having a natural hierarchy with multiple SOMs was presented in [3]. Such object hierarchies can be found e.g. in web pages, e-mail and MMS messages, and digital video. The multi-part hierarchy used for indexing video shots in this paper is illustrated in Fig. 1. The video shot itself is considered as the main or parent object. The keyframes (i.e. representative still images captured within the shot), audio track, and automatic speech recognition (ASR) text data are linked as children of the parent object. This hierarchy could also be extended further, e.g. the keyframe objects could have image segments as subobjects, the original full video is the video shot's parent, etc. All object modalities may have one or more SOMs or other feature indices, and thus all objects in the hierarchy may have links to a set of associated indices.

In this setting, the relevance of each object in the tree structure can be considered as a property of not only the object itself, but to some extent also of the other objects in the same structure. The ground-truth assessments are propagated from the parent, i.e. video shot, object to all children objects, which are then mapped to their corresponding SOMs, as described in more detail in Section 3. Finally, before deciding on the most likely shots associated with a semantic concept, the subobject scores are propagated back to the corresponding video shots.

2.2 Multimodal Features

In indexing video data with SOMs, we used in total four video features, six still image features, and one audio feature. A separate 256×256-sized SOM was trained for each of these eleven features. For the ASR text data, we used two alternative conceptwise text features based on an inverted file. These features are only briefly listed below, see [4] for more details.

Video features. On the video shot level, we used the MPEG-7 [5] *Motion Activity* (MA) descriptor and temporal versions of three still image features: *Average Color* (AC), *Color Moments* (CM) and *Texture Neighborhood* (TN). The temporal image features are calculated by dividing the shot into five equal parts and extracting averaged feature vectors for each part. The feature vector of the shot is then obtained by concatenating these five vectors.

Image features. For the keyframe indices we used a set of six standard MPEG-7 [5] descriptors, viz. *Color Layout* (CL), *Color Structure* (CS), *Dominant Color* (DC), *Scalable Color* (SC), *Edge Histogram* (EH), and *Homogeneous Texture* (HT). The descriptors were extracted globally from every keyframe in the collection, i.e. no segmentation or zoning was used.

Audio features. The Mel-scaled cepstral coefficient, or shortly *Mel Cepstrum* (CE) is the discrete cosine transform applied to the logarithm of the mel-scaled filter bank energies, appended with the total power of the signal.

Text features. Unlike the other features, an inverted file instead of a SOM index was used for the ASR output. The text features were constructed by gathering concept-dependent lists of 10 and 100 most informative terms.

3 Semantic Concepts as SOM Class Models

Assume that we have trained a SOM in an unsupervised fashion, using a large set of high-dimensional vectors. Let us choose a subset of vectors, which may be included in the original training set or be a new sample of similar data. The subset contains objects that are semantically related, as defined by a human user.

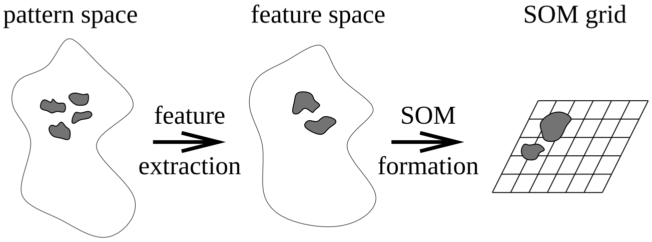

Fig. 2. Stages in creating a class model from the very-high-dimensional pattern space through the high-dimensional feature space to the two-dimensional SOM grid.

Such a subset is standardly *mapped* on the trained SOM by finding the BMU for each vector and counting the number of hits for each map unit. Normalized to unit sum, the hit frequencies give a discrete histogram which is a sample estimate of a probability distribution of the class on the SOM surface [6].

The shape of the distribution on the SOM surface depends on several factors:

- The distribution of the *original data* in the very-high-dimensional pattern space is generally given and cannot be controlled.
- The *feature extraction* technique in use affects the metrics and thus the distribution of all the generated feature vectors.
- The *overall shape* of the training set, after it has been mapped from the original data space to the feature vector space, determines the overall organization of the SOM.
- The *class distribution* of the studied object subset or class, relative to the overall shape of the feature vector distribution, specifies the layout of the class on the formed SOM.

Figure 2 visualizes how the pattern space is projected to feature space, the vectors of which are then used in training the SOM. The areas occupied by objects of a particular class are shown with gray shades.

In the very-high-dimensional pattern space the distribution of any non-trivial object class is most certainly sparse. As a consequence, in most cases it is meaningless to talk about the uni- or multimodality of class distributions in the pattern space. On the other hand, if the feature extraction stage is working properly, semantically similar patterns will in the feature space be mapped nearer to each other than semantically dissimilar ones. In the most advantageous situation, the pattern classes match clusters in the feature space, i.e. there exists a one-to-one correspondence between feature vector clusters and pattern classes. The relative distances between the feature vectors of a class compared to the overall distribution of the feature space data determine how well the class is concentrated on nearby SOM units. This can also be measured quantitatively [6].

Due to the topology preservation property of the SOM, one may now force the neighboring SOM units to interact by *low-pass filtering* or *convolving* the hit distributions on the SOM surface. When the surface is convolved, the one-to-one relationship between input vectors' SOM indices and hits on the SOM

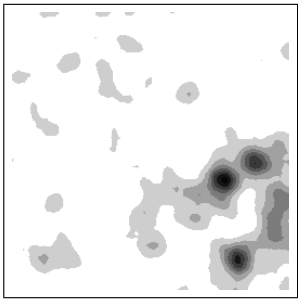

Fig. 3. An example class model (concept *explosion/fire* on the Color Layout SOM). Areas occupied by objects of the concept are shown with gray shades.

surface is broken. Instead, each hit results in a spread point response around the BMU. These class-conditional distributions or *class models* can be considered as estimates of the true distributions of the semantic concepts in question, not on the original feature spaces, but on the discrete two-dimensional grids defined by the used SOMs (see Fig. 3 for an example). Thereby, instead of modeling the probability density function in the high-dimensional feature spaces, we are essentially performing kernel-based estimation of class densities at the discrete distributions over the SOM surface. Depending on the variance of the kernel function, these kernels will overlap and weight vectors close to each other will partially share each other's probability mass.

For example, the most representative objects of a given semantic concept can be obtained by locating the SOM units, and the objects mapped to these units, that have the highest responses on the estimated class distribution. And, as the response values of the parallel indices are mutually comparable, we can determine a global ordering and the overall best candidate objects also when using multiple SOMs. By locating the corresponding objects in all SOM indices, we get their scores with respect to different features. The total scores for the candidate objects are then obtained by summing up the mapwise values. Furthermore, the shortcomings of different features with certain semantic concepts can be examined by studying the objects that yield a strong response on the class distributions but do not share the semantic content in question.

The responses invoked by different class models on the SOMs can also be directly used in automatic annotation of new objects. For this purpose, there are two distinct approaches. First, we can enumerate over all concepts and annotate those new objects that have the overall highest responses on the class models with the corresponding concept or annotation [7]. Alternatively, the input objects we want to annotate can be used to construct a new class distribution which is then compared to the existing models of semantic concepts using some distance measure suitable for probability distributions [8]. The latter approach is suited for the annotation of object groups sharing a semantic concept in a natural way; with more reference objects of a given concept available, the estimate of the corresponding distribution can be expected to become more accurate.

Table 1. Features used in the experiments for each concept to be detected.

semantic concept	video				image						audio	text	
	MA	AC	CM	TN	CL	CS	DC	SC	EH	HT	CE	10	100
walking/running	×	×							×				
explosion/fire			×		×					×			×
maps		×	×			×		×	×	×			×
flag-us	×		×	×					×			×	
building		×	×				×		×	×			×
waterscape/waterfront	×	×	×	×	×				×			×	
mountain	×	×		×					×	×	×		
prisoner										×			
sports	×	×	×		×				×			×	
car	×	×	×	×	×				×	×	×		×

4 Experiments

For associating specific semantic concepts with visual objects by using a generative approach, a method is needed for estimating the distribution of the concept over the feature representations of the training data. For this purpose, we use an existing lexicon for the development set of the TRECVID 2005 corpus and construct class models for the concepts to be detected (listed in Table 1; for the full definitions see [9, 10]), as described in Section 3. Thus, in these experiments, we do not use specialized detectors, but instead, all concepts are detected using the same procedure based on the ground-truth annotation of that concept.

The main video data for TRECVID 2005 [9] evaluations consists of about 170 hours of TV news in three languages (English, Chinese, Arabic) recorded in November 2004. In addition to the original videos transcoded to MPEG-1 format, a master shot reference [11], common keyframes for each shot, and automatic speech recognition output followed by automatic machine translation for the non-English news programs are provided. The data is split into development and test sets, with 43 907 and 45 766 shots in them, respectively. Furthermore, a joint effort to the participants to annotate the whole development set for 39 concepts (including the 10 concepts in the evaluation and 29 others) was organized. For this purpose, a downloadable tool for Windows platform provided by Carnegie Mellon University and a web-based tool [10] from IBM were available. In the end, most of the development set was in fact annotated twice, so we adopted a rule that a shot is considered relevant if either one of the annotators had accepted it.

In the high-level feature (concept) extraction task of TRECVID 2005, the purpose was to evaluate different detection methods for semantic concepts. Based on the annotation effort on the development set, the task was to return an ordered list of at most 2000 shots ranked according to the possibility of detecting the presence of the given concept in the shot. Due to the size of the test set, it was not evaluated in full for each concept. Instead, a pool of possibly relevant shots was first obtained by gathering sets of shots returned by the participating groups. These sets were then merged, duplicates removed, and the relevance of

Table 2. Detection results for each concept.

semantic concept	average precision			precision at depth			a priori prec.	
	PicSOM	median	max	100	1000	2000	devel.	test
walking/running	0.166	0.145	0.346	0.860	0.395	0.298	0.084	0.079
explosion/fire	0.026	0.037	0.129	0.160	0.055	0.037	0.010	0.009
maps	0.415	0.185	0.526	1.000	0.754	0.465	0.019	0.044
flag-us	0.064	0.071	0.253	0.280	0.091	0.065	0.007	0.011
building	0.226	0.236	0.511	0.970	0.465	0.350	0.073	0.076
waterscape/waterfront	0.344	0.187	0.493	0.970	0.340	0.218	0.026	0.019
mountain	0.305	0.155	0.458	0.920	0.282	0.180	0.013	0.016
prisoner	0.001	0.001	0.056	0.000	0.004	0.005	0.002	0.002
sports	0.210	0.231	0.521	0.560	0.234	0.143	0.040	0.013
car	0.200	0.181	0.369	0.960	0.441	0.297	0.067	0.045
mean	0.196	0.143	0.366	0.668	0.306	0.201	0.034	0.031

this subset is assessed manually. There were 22 participating groups submitting a total of 110 runs, all of which were pooled and judged to depth of 250 shots.

Instead of using a fixed set of features, we selected the set of used features for each concept separately. For this purpose, we applied a SFS-type feature selection scheme, in which we begin with an empty set and compute a criterion value for each of the potential features. If adding the feature with the highest value improves the overall result, that feature is added to the set of used features for that concept and the process is continued. Otherwise we stop the selection process. As the optimization criterion we used the average precision at 2000 returned items with two-fold cross validation on the development set.

The eleven features with SOM indices described in Section 2.2 along with the two concept-dependent text features were always included as potential features. The text features were alternative to each other, so only one of them could be selected. The conceptwise sets of selected features are listed in Table 1 (the feature abbreviations are listed in Section 2.2). As can be seen, the selection process typically resulted in 4–7 parallel features. The *prisoner* concept was a notable exception as adding any second feature, including the text features, beside Homogeneous Texture resulted in performance degradation.

The conceptwise results of detection performance are listed in Table 2. The (non-interpolated) average precision values are obtained by first determining and summing the precision at each location where a relevant shot is found and then dividing the result with the minimum of the total number of relevant shots or the maximum number of returned shots allowed (i.e. 2000). The maximum and median average precisions in Table 2 are also conceptwise, and do not therefore correspond to any single submission. The best single submission had a mean average precision of 0.336. It can be seen that the success of detecting different concepts varies considerably. Some concepts, such as *maps*, *building*, *waterscape/waterfront*, and *car* produce rather good results, especially in the beginning of the result list as can be seen from the "precision at depth 100" column, whereas detecting shots of the concept *prisoner* fails completely.

5 Conclusions

Statistical modeling of mid-level semantic concepts can be a very useful step in supporting high-level querying on visual data. In this paper, we described a method for applying multiple SOMs trained with multimodal features in semantic concept representation and detection. The class models for different semantic concepts were produced using a manually annotated video shot collection as the ground truth. For indexing video shots, we utilized a recently proposed method to support general hierarchical multimodal objects. The video shot, audio track, keyframes and ASR text data are all indexed separately and the ground-truth information and detection scores are propagated intrinsically.

The experiments reported in this paper were a part of our first time participation [4] in the annual TRECVID evaluation, and so we faced a lot of system development and other non-recurring work in order to be able to run the experiments. Therefore, we had limited time to study the effects of different setups and parameter values on the overall performance. Still, the results of the experiments are promising and can be seen to validate that SOM-based class models can be successfully used for detecting semantic concepts from multimodal data.

References

1. Kohonen, T.: Self-Organizing Maps. Third edn. Springer-Verlag (2001)
2. Laaksonen, J., Koskela, M., Oja, E.: PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing **13** (2002) 841–853
3. Sjöberg, M., Laaksonen, J.: Content-based retrieval of web pages and other hierarchical objects with Self-Organizing Maps. In: Proceedings of 15th International Conference on Artificial Neural Networks (ICANN 2005), Warsaw, Poland (2005)
4. Koskela, M., Laaksonen, J., Sjöberg, M., Muurinen, H.: PicSOM experiments in TRECVID 2005. In: TREC Video Retrieval Evaluation Online Proceedings, TRECVID (2005)
5. ISO/IEC: Information technology - Multimedia content description interface - Part 3: Visual (2002) 15938-3:2002(E).
6. Laaksonen, J., Koskela, M., Oja, E.: Class distributions on SOM surfaces for feature extraction and object retrieval. Neural Networks **17** (2004) 1121–1133
7. Viitaniemi, V., Laaksonen, J.: Keyword-detection approach to automatic image annotation. In: Proceedings of 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, London, UK (2005)
8. Koskela, M., Laaksonen, J.: Semantic annotation of image groups with Self-Organizing Maps. In: Proceedings of 4th International Conference on Image and Video Retrieval (CIVR 2005), Singapore (2005) 518–527
9. Over, P., Ianeva, T., Kraaij, W., Smeaton, A.F.: TRECVID 2005 - an introduction. In: TREC Video Retrieval Evaluation Online Proceedings, TRECVID (2005)
10. Volkmer, T., Smith, J.R., Natsev, A.P., Campbell, M., Naphade, M.: A web-based system for collaborative annotation of large image and video collections. In: Proc. 13th ACM International Conference on Multimedia, Singapore (2005)
11. Petersohn, C.: Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In: TREC Video Retrieval Evaluation Online Proceedings, TRECVID (2004)