Improving Automatic Video Retrieval with Semantic Concept Detection*

Markus Koskela, Mats Sjöberg, and Jorma Laaksonen

Department of Information and Computer Science,
Helsinki University of Technology (TKK), Espoo, Finland
{markus.koskela,mats.sjoberg,jorma.laaksonen}@tkk.fi
http://www.cis.hut.fi/projects/cbir/

Abstract. We study the usefulness of intermediate semantic concepts in bridging the semantic gap in automatic video retrieval. The results of a series of large-scale retrieval experiments, which combine text-based search, content-based retrieval, and concept-based retrieval, is presented. The experiments use the common video data and sets of queries from three successive TRECVID evaluations. By including concept detectors, we observe a consistent improvement on the search performance, despite the fact that the performance of the individual detectors is still often quite modest.

1 Introduction

Extracting semantic concepts from visual data has attracted a lot of attention recently in the field of multimedia analysis and retrieval. The aim of the research has been to facilitate semantic indexing of and concept-based retrieval from visual content. The leading principle has been to build semantic representations by extracting intermediate semantic levels (events, objects, locations, people, etc.) from low-level visual and aural features using machine learning techniques.

In early content-based image and video retrieval systems, the retrieval was usually based solely on querying by examples and measuring the similarity of the database objects (images, video shots) with *low-level features* automatically extracted from the objects. Generic low-level features are often, however, insufficient to discriminate content well on a conceptual level. This "semantic gap" is the fundamental problem in multimedia retrieval. The modeling of *mid-level semantic concepts* can be seen as an attempt to fill, or at least reduce, the semantic gap. Indeed, in recent studies it has been observed that, despite the fact that the accuracy of the concept detectors is far from perfect, they can be useful in supporting *high-level indexing and querying* on multimedia data [1]. This is mainly because such semantic concept detectors can be trained off-line with computationally more demanding algorithms and considerably more positive and negative examples than what are typically available at query time.

* Supported by the Academy of Finland in the *Finnish Centre of Excellence in Adaptive Informatics Research* project and by the TKK MIDE programme project UI-ART.

A.-B. Salberg, J.Y. Hardeberg, and R. Jenssen (Eds.): SCIA 2009, LNCS 5575, pp. 480–489, 2009.

In recent years, the TRECVID[1] [2] evaluations have emerged arguably as the leading venue for research on content-based video analysis and retrieval. TRECVID is an annual workshop series which encourages research in multimedia information retrieval by providing large test collections, uniform scoring procedures, and a forum for comparing results for participating organizations.

In this paper, we present a systematic study of the usefulness of semantic concept detectors in automatic video retrieval based on our experiments in three successive TRECVID workshops in the years 2006–2008. Overall, the experiments consist of 96 search topics with associated ground truth in test video corpora of 50–150 hours in duration. A portion of these experiments have been submitted to the official TRECVID evaluations, but due to the submission limitations in TRECVID, some of the presented experiments have been evaluated afterwards using the ground-truth provided by the TRECVID organizers.

The rest of the paper is organized as follows. Section 2 provides an overview of semantic concept detection and the method employed in our experiments. Section 3 discusses briefly the use of semantic concepts in automatic and interactive video retrieval. In Section 4, we present a series of large-scale experiments in automatic video retrieval, which combine text-based search, content-based retrieval, and concept-based retrieval. Conclusions are then given in Section 5.

2 Semantic Concept Detection

The detection and modeling of semantic mid-level concepts has emerged as a prevalent method to improve the accuracy of content-based multimedia retrieval. Recently published large-scale multimedia ontologies such as the Large Scale Concept Ontology for Multimedia (LSCOM) [3] as well as large annotated datasets (e.g. TRECVID, PASCAL Visual Object Classes[2], MIRFLICKR Image Collection[3]) have allowed an increase in multimedia concept lexicon sizes by orders of magnitude. As an example, Figure 1 lists and exemplifies the 36 semantic concepts detected for the TRECVID 2007 high-level feature extraction task. It should be elaborated that high-level feature extraction in TRECVID terminology corresponds to mid-level semantic concept detection.

Disregarding certain specific concepts for which specialized detectors exist (e.g. *human faces*, *speech*), the predominant approach to producing semantic concept detectors is to treat the problem as a generic learning problem, which makes it scalable to large ontologies. The concept-wise training data is used to learn independent detectors for the concepts over selected low-level feature distributions. For building such detectors, a popular approach is to use discriminative methods, such as SVMs, k-nearest neighbor classifiers, or decision trees, to classify between the positive and negative examples of a certain concept. In particular, SVM-based concept detection can be considered as the current *de facto* standard. The SVM detectors require, however, considerable computational resources for training the classifiers. Furthermore, the effect of varying background

[1] http://www-nlpir.nist.gov/projects/trecvid/
[2] http://pascallin.ecs.soton.ac.uk/challenges/VOC/
[3] http://press.liacs.nl/mirflickr/

Fig. 1. The set of 36 semantic concepts detected in TRECVID 2007

is often reduced by using local features such as the SIFT descriptors [4] extracted from a set of interest or corner points. Still, the current concept detectors tend to overfit to the idiosyncrasies of the training data, and their performance often drops considerably when applied to test data from a different source.

2.1 Concept Detection with Self-Organizing Maps

In the experiments reported in this paper, we take a generative approach in which the probability density function of a semantic concept is estimated based on existing training data using kernel density estimation. Only a brief overview is provided here; the proposed method is described in detail in [5].

A large set of low-level features is extracted from the video shots, keyframes extracted from the shots, and the audio track. Separate Self-Organizing Maps (SOMs) are first trained on each of these features to provide a common indexing structure across the different modalities. The positive examples in the training data for each concept are then mapped into the SOMs by finding the best matching unit for each example and inserting a local kernel function. These class-conditional distributions can then be considered as estimates of the true distributions of the semantic concepts in question—not on the original high-dimensional feature spaces, but on the discrete two-dimensional grids defined by the used SOMs. This reduction of dimensionality drastically reduces the computational requirements for building new concept models.

The particular feature-wise SOMs used for each concept detector are obtained by using some feature selection algorithm, e.g. sequential forward selection.

In the TRECVID high-level feature extraction experiments, the used approach has reached relatively good performance, although admittedly failing to reach the level of the current state-of-the-art detectors, which are usually based on SVM classifiers and thus require substantial computational resources for parameter optimization. Our method has, however, proven to be readily scalable to a large number of concepts, which has enabled us to model e.g. a total of 294 concepts from the LSCOM ontology and utilize these concept detectors in various TRECVID experiments without excessive computational requirements.

3 Concept-Based Video Retrieval

The objective of video retrieval is to find relevant video content for a specific information need of the user. The conventional approach has been to rely on textual descriptions, keywords, and other meta-data to achieve this functionality, but this requires manual annotation and does not usually scale well to large and dynamic video collections. In some applications, such as YouTube, the text-based approach works reasonably well, but it fails when there is no meta-data available or when the meta-data cannot adequately capture the essential content of the video material.

Content-based video retrieval, on the other hand, utilizes techniques from related research fields such as image and audio processing, computer vision, and machine learning, to automatically index the video material with low-level features (color layout, edge histogram, Gabor texture, SIFT features, etc.). Content-based queries are typically based on a small number of provided examples (i.e. *query-by-example*) and the database objects are rated based on their similarity to the examples according to the low-level features.

In recent works, the content-based techniques are commonly combined with separately pre-trained detectors for various semantic concepts (*query-by-concepts*) [6,1]. However, the use of concept detectors brings out a number of important research questions, including how to select the concepts to be detected, which methods to use when training the detectors, how to deal with the mixed performance of the detectors, how to combine and weight multiple concept detectors, and how to select the concepts used for a particular query instance.

Automatic Retrieval. In automatic concept-based video retrieval, the fundamental problem is how to map the user's information need into the space of available concepts in the used concept ontology [7]. The basic approach is to select a small number of concept detectors as active and weight them based either on the performance of the detectors or their estimated suitability for the current query. Negative or complementary concepts are not typically used.

In [7], Natsev et al. divide the methods for automatic selection of concepts into three categories: *text-based*, *visual-example-based*, and *results-based methods*. Text-based methods use lexical analysis of the textual query and resources such as WordNet [8] to map query words into concepts. Methods based on visual examples measure the similarity between the provided example objects and the concept detectors to identify suitable concepts. Results-based methods perform an initial retrieval step and analyze the results to determine the concepts that are then incorporated into the actual retrieval algorithm.

The second problem is how to fuse the output of the concept detectors with the other modalities such as text search and content-based retrieval. It has been observed that the relative performances of the modalities significantly depend on the types of queries [9,7]. For this reason, a common approach is to use *query-dependent fusion* where the queries are classified into one of a set of predetermined query classes (e.g. *named entity*, *scene query*, *event query*, *sports query*, etc.) and the weights for the modalities are set accordingly.

Interactive Retrieval. In addition to automatic retrieval, interactive methods constitute a parallel retrieval paradigm. Interactive video retrieval systems include the user in the loop at all stages of the retrieval session and therefore require sophisticated and flexible user interfaces. A global database visualization tool providing an overview of the database as well as a localized point-of-interest with increased level of detail are typically needed. Relevance feedback can also be used to manipulate the system toward video material the user considers relevant.

In recent works, semantic concept detection has been recognized as an important component also in interactive video retrieval [1], and current state-of-the-art interactive video retrieval systems (e.g. [10]) typically use concept detectors as a starting point for the interactive search functionality. A specific problem in concept-based interactive retrieval is how to present to a non-expert user the list of available concepts from a large and unfamiliar concept ontology.

4 Experiments

In this section, we present the results of our experiments in fully-automatic video search in the TRECVID evaluations of 2006–2008. The setup combines text-based search, content-based retrieval, and concept-based retrieval, in order to study the usefulness of existing semantic concept detectors in improving video retrieval performance.

4.1 TRECVID

The video material and the search topics used in these experiments are from the TRECVID evaluations [2] in 2006–2008. TRECVID is an annual workshop series organized by the National Institute of Standards and Technology (NIST), which provides the participating organizations large test collections, uniform scoring procedures, and a forum for comparing the results. Each year TRECVID contains a variable set of video analysis tasks such as high-level feature (i.e. concept) detection, video search, video summarization, and content-based copy detection. For video search, TRECVID specifies three modes of operation: fully-automatic, manual, and interactive search. Manual search refers to the situation where the user specifies the query and optionally sets some retrieval parameters based on the search topic before submitting the query to the retrieval system.

In 2006 the type of used video material was recorded broadcast TV news in English, Arabic, and Chinese, and in 2007 and 2008 the material consisted of documentaries, news reports, and educational programming from Dutch TV. The video data is always divided into separate development and test sets, with the amount of test data being approximately 150, 50, and 100 hours in 2006, 2007 and 2008, respectively. NIST also defines sets of standard search topics for the video search tasks and then evaluates the results submitted by the participants. The search topics contain a textual description along with a small number of both image and video examples of an information need. Figure 2 shows an example of a search topic, including a possible mapping of concept detectors from a concept

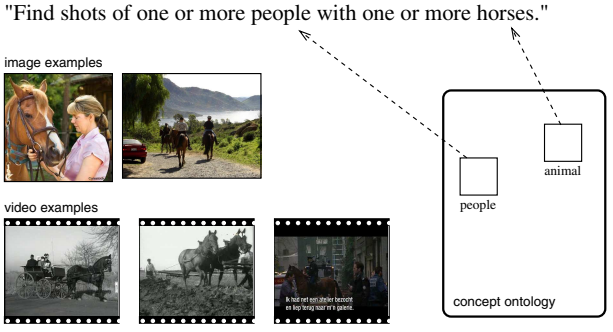

Fig. 2. An example TRECVID search topic, with one possible lexical concept mapping from a concept ontology

ontology based on the textual description. The number of topics evaluated for automatic search was 24 for both 2006 and 2007 and 48 for the year 2008. Due to the limited space, the search topics are not listed here, but are available in the TRECVID guidelines documents at *http://www-nlpir.nist.gov/projects/trecvid/*

The video material used in the search tasks is divided into shots in advance and these reference shots are used as the unit of retrieval. The output from an automatic speech recognition (ASR) software is provided to all participants. In addition, the ASR result from all non-English material is translated into English by using automatic machine translation.

Due to the size of the test corpora, it is infeasible within the resources of the TRECVID initiative to perform an exhaustive examination in order to determine the topic-wise ground truth. Therefore, the following pooling technique is used instead. First, a pool of possibly relevant shots is obtained by gathering the sets of shots returned by the participating teams. These sets are then merged, duplicate shots are removed, and the relevance of only this subset of shots is assessed manually. It should be noted that the pooling technique can result in the underestimation of the performance of new algorithms and, to a lesser degree, new runs, which were not part of the official evaluation, as all unique relevant shots retrieved by them will be missing from the ground truth.

The basic performance measure in TRECVID is *average precision* (AP):

$$\mathrm{AP} = \frac{\sum_{r=1}^{N}(P(r) \times R(r))}{N_{rel}} \tag{1}$$

where r is the rank, N is the number of retrieved shots, $R(r)$ is a binary function stating the relevance of the shot retrieved with rank r, $P(r)$ is the precision at the rank r, and N_{rel} is the total number of relevant shots in the test set. In TRECVID search tasks, N is set to 1000. The mean of the average precision values over a set of queries, *mean average precision* (MAP) has been the standard evaluation measure in TRECVID. In recent years, however, average precision has been gradually replaced by *inferred average precision* (IAP) [11], which approximates the AP measure very closely but requires only a subset of the pooled results

to be evaluated manually. The query-wise IAP values are similarly combined to form the performance measure *mean inferred average precision* (MIAP).

4.2 Settings for the Retrieval Experiments

The task of automatic search in TRECVID has remained fairly constant over the three year period in question. Our annual submissions have been, however, somewhat different each year due to modifications and additions to our PicSOM [12] retrieval system framework, to the used features and algorithms, etc. For brevity, only a general overview of the experiments and the used settings is provided in this paper. More detailed descriptions can be found in our annual TRECVID workshop papers [13,14,15]. In all experiments, we combine content-based retrieval based on the topic-wise image and video examples using our standard SOM-based retrieval algorithm [12], concept-based retrieval with concept detectors trained as described in Section 2.1, and text search (c.f. Fig. 2).

The semantic concepts are mapped to the search topics using lexical analysis and synonym lists for the concepts obtained from WordNet. In 2006, we used a total of 430 semantic concepts from the LSCOM ontology. However, the LSCOM ontology is currently annotated only for the TRECVID 2005/2006 training data. Therefore, in 2007 and 2008, we used only the concept detectors available from the corresponding high-level feature extraction tasks, resulting in 36 and 53 concept detectors, respectively. In the 2008 experiments, 11 of the 48 search topics did not match to any of the available concepts. The visual examples were used instead for these topics.

For text search, we employed our own implementation of an inverted file index in 2006. For the 2007–2008 experiments, we replaced our indexing algorithm with the freely-available Apache Lucene[4] text search engine.

4.3 Results

The retrieval results for the three studied TRECVID test setups are shown in Figures 3–5. The three leftmost (lighter gray) bars show the retrieval performance of each of the single modalities: text search ('t'), content-based retrieval based on the visual examples ('v'), and retrieval based on the semantic concepts ('c'). The darker gray bars on the right show the retrieval performances of the combinations of the modalities. The median values for all submitted comparable runs from all participants are also shown as horizontal lines for comparison.

For 2006 and 2007, the shown performance measure is mean average precision (MAP), whereas in 2008 the TRECVID results are measured using mean inferred average precision (MIAP). Direct numerical comparison between different years of participation is not very informative, since the difficulty of the search tasks may vary greatly from year to year. Furthermore, the source of video data used was changed between years 2006 and 2007. Relative changes, however, and changes between different types of modalities can be very instructive.

[4] http://lucene.apache.org

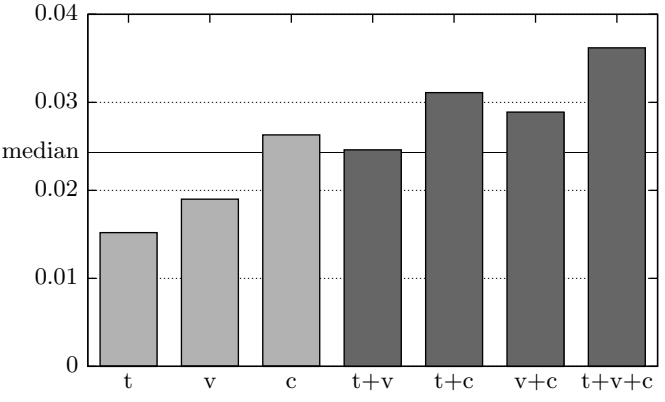

Fig. 3. MAP values for TRECVID 2006 experiments

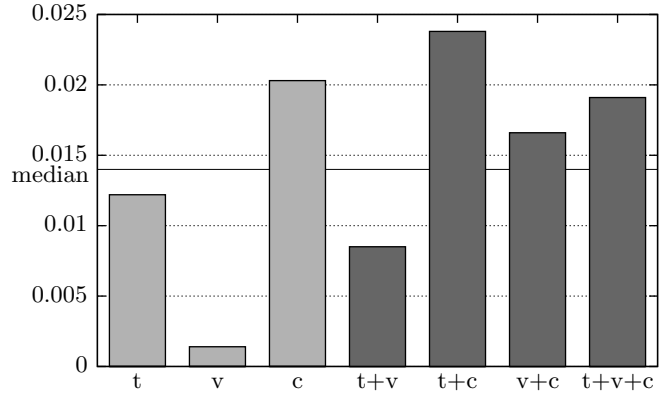

Fig. 4. MAP values for TRECVID 2007 experiments

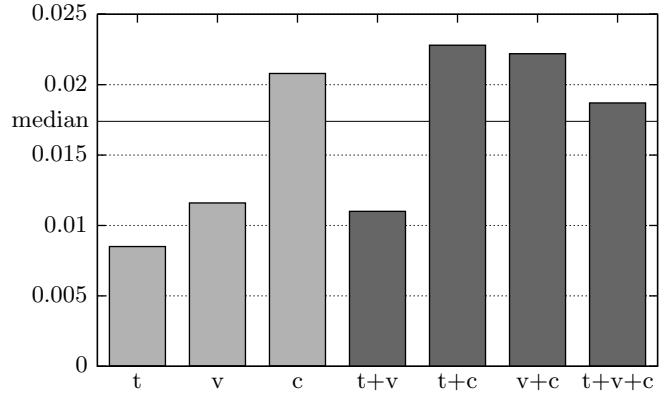

Fig. 5. MIAP values for TRECVID 2008 experiments

The good relative performance of the semantic concepts can be readily observed from Figures 3–5. In all three sets of single modality experiments, the concept-based retrieval has the highest performance. Content-based retrieval, on the other hand, shows considerably more variance in performance, especially when considering the topic-wise AP/IAP results (not shown due to space limitations) instead of the mean values considered here. In particular, the visual examples in the 2007 runs seem to perform remarkably modestly. This can be readily explained by examining the topic-wise results: It turns out that most of the content-based results are indeed quite poor, but in 2006 and 2008 there were a few visual topics for which the visual features were very useful.

A noteworthy aspect in the TRECVID search experiments is the relatively poor performance of text-based search. This is a direct consequence of both the low number of named entity queries among the search topics and the noisy text transcript resulting from automatic speech recognition and machine translation.

Of the combined runs, the combination of text search and concept-based retrieval performs reasonably well, resulting in the best overall performance in the 2007 and 2008 and second-best results in the 2006 experiments. Moreover, it reaches better performance than any of the single modalities in all three experiment setups. Another way of examining the results of the experiments is to compare the runs where the concept detectors are used with the corresponding ones without the detectors (i.e. 't' vs 't+c', 'v' vs 'v+c' and 't+v' vs 't+v+c'). Viewed this way, we observe a strong increase in performance in all cases by including the concept detectors.

5 Conclusions

The construction of visual concept lexicons or ontologies has been found to be an integral part of any effective content-based multimedia retrieval system in a multitude of recent research studies. Yet the design and construction of multimedia ontologies still remains an open research question. Currently the specification of which semantic features are to be modeled tends to be fixed irrespective of their practical applicability. This means that the set of concepts in an ontology may be appealing from a taxonomic perspective, but may contain concepts which make little difference in their discriminative power.

The appropriate use of the concept detectors in various retrieval settings is still another open research question. Interactive systems—with the user in the loop—require solutions different from those used in automatic retrieval algorithms which cannot rely on human knowledge in the selection and weighting of the concept detectors.

In this paper, we have presented a comprehensive set of retrieval experiments with large real-world video corpora. The results validate the observation that semantic concept detectors can be a considerable asset in automatic video retrieval, at least with the high-quality produced TV programs and TRECVID style search topics used in these experiments. This holds even though the performance of the individual detectors is inconsistent and still quite modest in

many cases, and though the mapping of concepts to search queries was performed using a relatively naïve lexical matching approach. Similar results have been obtained in the other participants' submissions to the TRECVID search tasks as well. These findings strengthen the notion that mid-level semantic concepts provide a true stepping stone from low-level features to high-level human concepts in multimedia retrieval.

References

1. Hauptmann, A.G., Christel, M.G., Yan, R.: Video retrieval based on semantic concepts. Proceedings of the IEEE 96(4), 602–622 (2008)
2. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: MIR 2006: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321–330. ACM Press, New York (2006)
3. Naphade, M., Smith, J.R., Tešić, J., Chang, S.F., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J.: Large-scale concept ontology for multimedia. IEEE MultiMedia 13(3), 86–91 (2006)
4. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
5. Koskela, M., Laaksonen, J.: Semantic concept detection from news videos with self-organizing maps. In: Proceedings of 3rd IFIP Conference on Artificial Intelligence Applications and Innovations, Athens, Greece, June 2006, pp. 591–599 (2006)
6. Snoek, C.G.M., Worring, M.: Are concept detector lexicons effective for video search? In: Proceedings of the IEEE International Conference on Multimedia & Expo. (ICME 2007), Beijing, China, July 2007, pp. 1966–1969 (2007)
7. Natsev, A.P., Haubold, A., Tešić, J., Xie, L., Yan, R.: Semantic concept-based query expansion and re-ranking for multimedia retrieval. In: Proceedings of ACM Multimedia (ACM MM 2007), Augsburg, Germany, September 2007, pp. 991–1000 (2007)
8. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database
9. Kennedy, L.S., Natsev, A.P., Chang, S.F.: Automatic discovery of query-class-dependent models for multimodal search. In: Proceedings of ACM Multimedia (ACM MM 2005), Singapore, November 2005, pp. 882–891 (2005)
10. de Rooij, O., Snoek, C.G.M., Worring, M.: Balancing thread based navigation for targeted video search. In: Proceedings of the International Conference on Image and Video Retrieval (CIVR 2008), Niagara Falls, Canada, pp. 485–494 (2008)
11. Yilmaz, E., Aslam, J.A.: 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)
12. 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(4), 841–853 (2002)
13. Sjöberg, M., Muurinen, H., Laaksonen, J., Koskela, M.: PicSOM experiments in TRECVID 2006. In: Proceedings of the TRECVID 2006 Workshop, Gaithersburg, MD, USA (November 2006)
14. Koskela, M., Sjöberg, M., Viitaniemi, V., Laaksonen, J., Prentis, P.: PicSOM experiments in TRECVID 2007. In: Proceedings of the TRECVID 2007 Workshop, Gaithersburg, MD, USA (November 2007)
15. Koskela, M., Sjöberg, M., Viitaniemi, V., Laaksonen, J.: PicSOM experiments in TRECVID 2008. In: Proceedings of the TRECVID 2008 Workshop, Gaithersburg, MD, USA (November 2008)