Exploiting Noisy Visual Concept Detection to Improve Spoken Content based Video Retrieval

Stevan Rudinac, Martha Larson, Alan Hanjalic
Multimedia Information Retrieval Lab, Delft University of Technology, Delft, The Netherlands
{s.rudinac, m.a.larson, a.hanjalic}@tudelft.nl

ABSTRACT
In this paper, we present a technique for unsupervised construction of concept vectors, concept-based representations of complete video units, from the noisy shot-level output of a set of visual concept detectors. We deploy these vectors to improve spoken-content-based video retrieval using Query Expansion Selection (QES). Our QES approach analyzes results lists returned in response to several alternative query expansions, applying a coherence indicator calculated on top-ranked items to choose the appropriate expansion. The approach is data driven, does not require prior training and relies solely on the analysis of the collection being queried and the results lists produced for the given query text. The experiments, performed on two datasets, TRECVID 2007/2008 and TRECVID 2009, demonstrate the effectiveness of our approach and show that a small set of well-selected visual concept detectors is sufficient to improve retrieval performance.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Retrieval models
General Terms
Algorithms, Performance, Experimentation

Keywords
Spoken Content Retrieval, Video-level Retrieval, Query Performance Prediction, Query Expansion Selection, Visual Concept Detectors, Semantic-Theme-based Video Retrieval

1. INTRODUCTION
In this paper, we present an unsupervised technique for creating concept vectors, video-level representations that are based on the shot-level output of visual concept detectors. Instead of using concept vectors to directly represent the semantic content of videos, we use them only to compare videos. The use of concept vector comparisons is based on the insight that the patterns of the visual concepts present in a video can capture resemblance between videos with respect to semantic theme, i.e., the overall topic of the video. Our previous work [13] has shown that a coherence-based framework for Query Expansion Selection (QES) can successfully improve video-level semantic-theme-based retrieval by making use of comparisons between concept vectors. In our QES framework, the choice between several alternative query expansions is made on the basis of the coherence indicators calculated using top-ranked videos in the results lists returned by a spoken-content-based retrieval system in response to each expansion. By using concept vectors to calculate the coherence indicator, we create an approach that successfully improves the output of a spoken-content-based video retrieval system by exploiting noisy information from visual concept detectors. The method has the advantages of avoiding the need for a mapping between queries and concepts, a technique that has been studied for shot-level retrieval, cf. [6], and of preventing undue sensitivity to concept noise. We do not assume that a special set of visual concepts must be detected for a given video collection. In other words, our approach does not require an assurance that the concept set used provides complete semantic coverage of the visual content of the collection. In this paper, we focus on concept selection and address the issue: Given the output of a set of visual concept detectors, how can we choose, in an unsupervised fashion, which concepts to include in the visual concept vector such that it can be used to calculate a coherence score that will improve retrieval performance within the QES framework.

In the next section, we briefly review related work. Then, we present our method for building concept vectors using unsupervised concept selection. We go on to present the experimental framework, including details of our QES framework and data sets. Finally, we report on experiments and discuss the results.

2. RELATED WORK
Query Expansion is widely used in information retrieval to enrich the original query (i.e., add query terms) so as to provide a better match with documents in the target collection. In the area of spoken content retrieval, query expansion is often used [8] where it also compensates for errors in the speech recognition transcripts, a known problem in spoken content retrieval. The danger of query expansion is that it introduces inappropriate terms into the query, causing topical drift. Topical drift can be controlled by appropriate query performance prediction applied to select which queries should be expanded [3]. In particular, our work is related to methods for post-retrieval query prediction, i.e., methods that use results lists returned by an initial retrieval run as the basis for their performance prediction [2, 18]. More recently, a computationally inexpensive coherence-based approach to query difficulty prediction has been proposed [5]. This approach measures the topical coherence of top documents in the results list. In our recent work [12, 13], we demonstrated the performance of the coherence score applied to query expansion selection for spoken content retrieval.

An important difference between the work presented here and previous work is the type of retrieval task. Semantic theme-based retrieval involves retrieving videos according to subject matter and the typical semantic theme queries are thus very different
from conventional TRECVID queries, which include named persons, named objects, general objects, scenes and sports. TRECVID-type queries are strongly related to the visual channel and may not be actually representative of the overall topic of the video. As mentioned above, this difference is reflected in the size of the retrieval unit. Unlike the majority of approaches that address video retrieval at the shot level (e.g., [15, 11]) we consider entire videos as retrieval units. Recently, [1] investigated video retrieval beyond the shot level, but with respect to a task in which relevance was judged based on visual content rather than semantic theme.

3. CONCEPT VECTORS

3.1 Building Concept Vectors

A critical challenge to making effective use of visual concepts is the relatively low performance of state-of-the-art visual concept detectors. As an example, the performance in terms of Mean Average Precision (MAP) of the best performer in “Concept Detection” and “Interactive Search” tasks of TRECVID 2009 was below 0.25 in those tasks [15]. Our approach is based on the insight that in spite of a relatively poor performance and noisiness of visual concept detectors at the shot level, aggregating the results of concept detections across a series of shots could still provide the basis for a representation that would be useful for computing video-level similarity.

Our concept vectors represent individual videos and contain one component for each concept \( c \) in the concept set \( C \). We refer to concept weights as term frequency (TF) weights to emphasize the similarity of the concept vector with document vectors used in text retrieval. The concept vector \( F_c \) for video \( v \) can be defined as,

\[
F_c = \{T_F(c), c \in C\}
\]  

(1)

The analogy is not perfect since concept detectors produce scores rather than counts [7, 14]. We calculate \( T_F(c) \) as the sum of confidence scores \( \text{Conf}_c \) of concept \( c \) over all shots \( j \) in video \( v \), normalized by the number of shots in \( v \).

\[
T_F(c) = \frac{\sum_j \text{Conf}_c(j)}{N_v}
\]  

(2)

We explored extending the analogy with text retrieval and making use of an inverse document frequency (IDF) component to the concept weight [13]. IDF introduced no additional benefit, and is not used further here.

An analysis of the outputs of state-of-the-art concept detectors (i.e., [14] and [7]) revealed that the values of visual concept confidence vary widely within the interval of [0, 1]. In order to prevent concept outputs with very low values from introducing noise into the system, we discard shot-level confidences whose scores do not exceed a specific threshold. Preliminary experiments indicated that the exact value of this threshold is not critical for overall performance. For the purposes of the experiments reported here, we set the threshold to 0.5.

3.2 Concept Selection

The goal of concept selection is to choose a set of concepts that is able to cleanly capture semantic similarities between videos. We assume that the suitability of semantic concepts in this regard is dependent on the characteristics of a particular video collection. We would like our concept selection approach to yield explicit information about which concepts are most valuable for a particular collection and for this reason do not pursue approaches that reduce the feature space, since they obscure the original dimensions making the primary contribution. Feature selection is a well-known problem in the pattern recognition and information retrieval communities and through years many efficient solutions were proposed [16, 17]. In contrast to most existing feature selection approaches, our approach is required to be completely data driven and unsupervised. We start from the assumption that a small set of well selected features might improve results significantly [16] and propose a simple unsupervised heuristic to select a subset of particularly informative concepts for a given video collection. This selection is based on concept frequency, variance and kurtosis across the collection.

Frequency: We conjecture that concepts that occur in many videos within the collection will be helpful in comparing videos. The relative difference in the confidence scores provides a basis for calculating similarity.

Variance: It is not only enough to select the frequent concepts, however, since some frequent concepts might have a confidence value that is completely uniform throughout the collection. Therefore, we would like these frequent concepts to have a high variance as well.

Kurtosis: A high variance might be the consequence of either infrequent extreme deviations or preferably frequent, but moderate variations of concept weights. To isolate the concepts with frequent but moderate variations in the given collection, we focus on those concepts with a low kurtosis.

Our approach builds a set of concepts that are simultaneously optimal with respect to all three measures. In order to determine that set, we produce three ranked lists by sorting the concepts according to the decreasing frequency and variance and increasing kurtosis in the collection. We then compute the percentage of concepts common to the top-N concepts from each list. We gradually increase \( N \) until this percentage reaches its first dominant peak (e.g., overlap of more than 70%) and select the concepts that are common to all three lists.

4. EXPERIMENTAL FRAMEWORK

4.1 Visual Concept based Coherence for QES

Our QES approach selects between results lists returned by spoken-content retrieval in response to the original query and multiple query expansions. The best result list is returned as the final result. Effectively QES makes a decision about whether the query should be expanded and if so, which of the alternative expansions should be used for optimal results. In the experiments presented here, the query is expanded using the following expansions a) Conventional PRF (pseudo-relevance feedback), b) WordNet and c) Google Sets expansion, in which we limit the number of expansion items (words or multword phrases) to 15.

A coherence indicator [5] is used to select the results list with the highest coherence among the top-N retrieved results. The indicator is computed according to Eq. (3).

\[
\text{Co}(\text{TopN}) = \sum_{i,j \in (\text{TopN})} \delta(v_i, v_j) \cdot \delta(v_i, v_j) \cdot \left\{ \begin{array}{ll}
1, & \text{sim}(v_i, v_j) > \theta \\
0, & \text{otherwise}
\end{array} \right.
\]  

(3)

In the equation above, the videos, \( v_i \) are represented using the concept vector \( F_c \). The threshold \( \theta \) is a similarity value between particularly close videos in the overall video collection. For our visual-concept-based coherence approach to QES the similarity is calculated using the concept vectors. As a similarity measure \( \text{sim} \) we use the cosine similarity.
4.2 Data and Semantic Concept Detectors

The experiments are performed on the MediaEval 2010 [9] Tagging Task Professional data sets, which are re-issues of the TRECVID 2007, 2008 and 2009 data sets from The Netherlands Institute for Sound and Vision (S&V). The data set contains Dutch-language television content, including news reports and magazines, as well as documentaries and other educational programming. MediaEval provided ground truth in the form of semantic theme labels that were assigned by professional archivists at S&V. Please note that a few videos for which no metadata were available were removed, making these data sets smaller than the original TRECVID versions. The MediaEval issue of TRECVID 07+08 contains 405 videos and 37 semantic labels used as queries. The MediaEval issue of TRECVID 09 contains 378 videos and 41 semantic theme labels.

We build concept vectors using the CU-VIREO374 concept detection scores [7], the output of detectors trained for 374 visual concepts selected from the LSCOM ontology [10]. For TRECVID 09, we also make use of MediaMill concept detector output [14] (64 concepts).

5. Experimental Results

We perform experiments to examine the usability of our concept-based video representation for semantic-theme-based video retrieval and to determine if the proposed concept selection approach improves the quality of query expansion selection. We confirm our result by testing on multiple data sets and investigating sensitivity to concept detector quality and parameter settings.

On both TRECVID 07+08 and TRECVID 09 we compare our method with the simple text-search baseline and three additional results lists produced using common query expansions. The performance of these approaches expressed in terms of the mean average precision (MAP) is shown in Table 1. Note that these four results lists are the ones that will be combined by our QES approach.

Table 1. MAP of the baseline and the query expansions used

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Baseline</th>
<th>PRF</th>
<th>WordNet</th>
<th>Google Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRECVID 07+08</td>
<td>0.2322</td>
<td>0.2619</td>
<td>0.1941</td>
<td>0.1271</td>
</tr>
<tr>
<td>TRECVID 09</td>
<td>0.2381</td>
<td>0.2621</td>
<td>0.1867</td>
<td>0.1276</td>
</tr>
</tbody>
</table>

When interpreting the results, it is important to note that here we are not interested in improving the MAP in the absolute sense since MAP depends on the quality of the baseline results lists. Instead, for each query, we target to always choose the best results list, whatever MAP it has. In order to make a comparison with the theoretical optimum, we make use of “oracle” indicators, hypothetical indicators that always choose the correct query expansion.

5.1 Without Concept Selection

In our first query expansion selection (QES) experiment, we used all 374 CU-VIREO374 and all 64 MediaMill concept detectors. In Table 2 performance for optimal parameter settings is reported along with the statistical significance with respect to the best condition in Table 1 (which we refer to as the “best baseline”).

For these experiments and all the experiments that follow, results are reported using optimal values of the parameters involved in calculating the coherence score in Eq. 5 ($\theta$ and the number of top-N items used).

Table 2. MAP of our QES approach when all concepts are used; statistically significant improvement over the best baseline is indicated with “*” (Wilcoxon Signed Rank test, p<0.05)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Concepts</th>
<th>Best base.</th>
<th>QES</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRECVID 07+08</td>
<td>CU-VIREO374</td>
<td>0.2619</td>
<td>0.233</td>
<td>0.3017</td>
</tr>
<tr>
<td>TRECVID 09</td>
<td>CU-VIREO374</td>
<td>0.2621</td>
<td>0.2588</td>
<td>0.3136</td>
</tr>
<tr>
<td>TRECVID 09</td>
<td>MediaMill 64</td>
<td>0.2621</td>
<td>0.2743*</td>
<td>0.3136</td>
</tr>
</tbody>
</table>

TRECVID 07+08 does not benefit from QES when a feature vector including all concepts is used to calculate coherence. For TRECVID 09, improvement is observed when MediaMill concept detector outputs are used, but is limited to specific parameter settings and is thus not robust. These experiments support our hypothesis that concept selection is a key step.

5.2 Applying Concept Selection

Next, we investigate the performance improvement that can be gained with concept selection. In Table 3, it can be seen that, although the results are still relatively far from the ideal performance of the oracle, QES with concept selection succeeds in consistently improving performance over the best baseline. Further, when using CU-VIREO374 performance improves over QES without concept selection (cf. Table 2).

Table 3. MAP of QES approach when our concept selection approach is used; statistically significant improvement over the best baseline marked by “*” (Wilcoxon Signed Rank test, p<0.05)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Concepts</th>
<th>Best base.</th>
<th>QES</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRECVID 07+08</td>
<td>CU-VIREO374</td>
<td>0.2619</td>
<td>0.2757*</td>
<td>0.3017</td>
</tr>
<tr>
<td>TRECVID 09</td>
<td>CU-VIREO374</td>
<td>0.2621</td>
<td>0.2631</td>
<td>0.3136</td>
</tr>
<tr>
<td>TRECVID 09</td>
<td>MediaMill 64</td>
<td>0.2621</td>
<td>0.2688*</td>
<td>0.3136</td>
</tr>
</tbody>
</table>

When using CU-VIREO374 concepts (374 concepts in total), our method selected 15 concepts for TRECVID 07+08 and 32 concepts for TRECVID 09. When using MediaMill concepts (64 concepts in total), our method selected 14 concepts. Note that the MAP that QES achieves using the reduced set of 14 MediaMill concepts is quite close to what is achieved with the full set. Although the selected set of concepts does not outperform the entire set in the case of MediaMill concepts, this result is nonetheless interesting since it shows that a small set of well-selected concepts can achieve improvement in video retrieval. The selected MediaMill concepts were: Building, Crowd, Face, Hand, Outdoor, Person, PersonWalkingOrRunning, Road, Sky, Street, TwoPeople, Urban, Vegetation and Waterscape. The members of this set are related to people or general scenes and not related to specific objects (i.e., Airplane, Chair and Flower were discarded). This difference may possibly have implications for decisions concerning the development of visual concept detectors.

We would like to note that these results suggest that there is a relationship between the quality of the concept detectors and their usefulness for our approach. MediaMill system achieved the highest performance in, e.g., TRECVID 2009 concept detection and interactive search tasks [15] and this concept set achieves better performance in our setting as well.
5.3 Robustness of Concept Vectors
In order to understand whether our concept selection approach is capable of selecting the optimal cutoff, we perform an additional experiment using TRECVID 07+08 and the CU-VIREO374 concept set. We gradually increase the number of top-N concepts from the lists of concepts ranked by frequency, variance and kurtosis. Increasing N means increasing the number of concepts selected. The best overall MAP of 0.2757 is obtained when 15 concepts are selected. The same set is chosen by our automatically selected cutoff. This result confirms that our approach is capable of choosing a cutoff optimal for building useful concept vectors.

To investigate the robustness of the coherence indicator to parameter setting, we investigate the performance of the system for several values of threshold θ (cf. Table 4). The number of items used to calculate the indicator is set to 20.

Table 4. MAP for different values of parameter θ; statistically significant improvement over the best query expansion method is indicated with ‘^’ (Wilcoxon Signed Rank test, p<0.05)

<table>
<thead>
<tr>
<th>θ (%)</th>
<th>TRECVID 07+08</th>
</tr>
</thead>
<tbody>
<tr>
<td>70%</td>
<td>0.2735^</td>
</tr>
<tr>
<td>80%</td>
<td>0.2745^</td>
</tr>
<tr>
<td>90%</td>
<td>0.2757^</td>
</tr>
<tr>
<td></td>
<td>0.2619</td>
</tr>
</tbody>
</table>

These results demonstrate the stability of the improvement in MAP achieved by our approach across a range of θ values.

6. Summary and Conclusion
Our approach for unsupervised construction of concept vectors was validated by a series of experiments that demonstrated improved retrieval performance within our QES framework for spoken-content-based video retrieval. The method of automatic concept selection that we proposed was shown to be transferable in an unproblematic manner to an unseen data set (TRECVID 09). The approach appears to be sensitive to the quality of the concept detector output, with better performing concept detectors yielding greater improvement. Note that changing data sets requires a re-optimization of the parameters involved in calculating the coherence indicator (θ and the number of top-N items used). However, additional tests showed the robustness of the approach and in particular the stability of results across settings for θ.

Our automatic approach for creating video-level representations builds concept vectors involving a relatively small number of concepts. If a small set of concept detectors is enough to improve the results of semantic-theme-based video retrieval, a productive avenue for concept detector research is to concentrate on achieving high quality for a small set of well-selected detectors and not on training concept detectors that will cover the entire conceivable semantic space.

Future work will involve investigation into the further refinement of our approach for building concept-based video level representations. In particular, we are interested in exploiting not only the frequency of occurrence of concepts, but also detailed information about their occurrence patterns, including distributional properties such as burstiness and also co-occurrence with other concepts. Finally, we are interested in investigating methods for automatically estimating the optimal parameter settings for QES, in determining the lower bound of concept detection quality necessary for a concept detector to be useful in our method and also determining the exact nature of the collection specific properties that make our approach more or less suitable for a particular semantic-theme-based retrieval task.

7. ACKNOWLEDGMENTS
The research leading to these results has received funding from the European Commission’s 7th Framework Programme (FP7) under grant agreement n° 216444 (NoE PetaMedia).

8. REFERENCES