Int J Multimed Info Retr (2012) 1:263–280
DOI 10.1007/s13735-012-0018-0

REGULAR PAPER

Leveraging visual concepts and query performance prediction for semantic-theme-based video retrieval

Stevan Rudinac · Martha Larson · Alan Hanjalic

Received: 14 February 2012 / Revised: 8 June 2012 / Accepted: 27 July 2012 / Published online: 24 September 2012

Abstract In this paper, we present a novel approach that utilizes noisy shot-level visual concept detection to improve text-based video retrieval. As opposed to most of the related work in the field, we consider entire videos as the retrieval units and focus on queries that address a general subject matter (semantic theme) of a video. Retrieval is performed using a coherence-based query performance prediction framework. In this framework, we make use of video representations derived from the visual concepts detected in videos to select the best possible search result given the query, video collection, available search mechanisms and the resources for query modification. In addition to investigating the potential of this approach to outperform typical text-based video retrieval baselines, we also explore the possibility to achieve further improvement in retrieval performance through combining our concept-based query performance indicators with the indicators utilizing the spoken content of the videos. The proposed retrieval approach is data driven, requires no prior training and relies exclusively on the analyses of the video collection and different results lists returned for the given query text. The experiments are performed on the Media Eval 2010 datasets and demonstrate the effectiveness of our approach.

S. Rudinac (✉) · M. Larson · A. Hanjalic
Multimedia Information Retrieval Lab, Delft University of Technology, Delft, The Netherlands
e-mail: s.rudinac@tudelft.nl

M. Larson
e-mail: m.a.larson@tudelft.nl

A. Hanjalic
e-mail: a.hanjalic@tudelft.nl

Keywords Semantic-theme-based video retrieval · Query performance prediction · Query expansion selection · Visual concept selection

1 Introduction

In this paper, we address the problem of video retrieval at the *semantic theme* level, where semantic theme refers to a general subject matter (topic) of a video. The query is formulated to encode a topical information need of the user, and the retrieval system is expected to return videos that treat relevant subjects. Examples of such "topical" queries are *court hearings*, *youth programs*, *archaeology*, *celebrations*, *scientific research*, *economics*, *politics* and *zoos*.

Semantic themes come in a variety of abstraction levels and degrees to which they are visually constraining. In practice, a set of semantic themes might include video genres in a more traditional sense [2,32] or the semantic labels assigned by archivists in professional libraries. They can, however, also correspond to the categories used in online content sharing portals, such as YouTube[1] and blip.tv[2].

A high level of inter-annotator agreement observed in professional digital libraries indicates that humans easily agree on the semantic theme of a video. Although it is not obvious where this inter-annotator agreement comes from, we hypothesize that both the visual and spoken content channel (ASR output) provide valuable information in this respect. While support for this hypothesis in the case of the spoken content channel was provided in our previous work [23], our goal in this paper is to investigate the potential of the visual channel to help retrieve videos using topical queries.

[1] http://www.youtube.com.

[2] http://www.blip.tv.

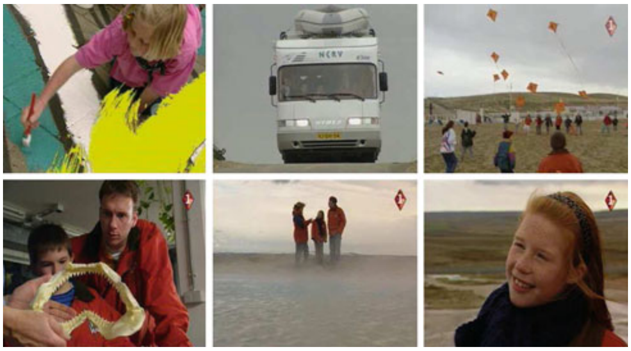


Fig. 1 Keyframes of shots from a video in the TRECVID 2009 collection that is relevant to the semantic theme *youth programs*. The visual content of the shots contains information only weakly related to what the entire video is actually about

On a first sight, the information in the visual channel may seem rather unreliable as an indicator of the general topic of a video. As shown in the examples in Fig. 1, frames extracted from different shots of a video covering the topic *youth programs* are characterized by highly diverse visual content that also does not directly connect a shot to the topic specified by the query. However, in view of the fact that the visual channel is used to complement or illustrate the topic of a video, it should not be surprising if the same key elements of the visual content, such as objects or parts of the scenery, appear in a large number of video clips covering the same semantic theme. Observed from this perspective, the visual content across different video shots in Fig. 1 may indeed be found consistent at a particular level of content representation, namely at the level of visual concepts. Here, our definition of a visual concept corresponds to the definition adopted in the TRECVID benchmark [19] and represented by the ontologies such as the LSCOM [17]. Typical examples of visual concepts are *vehicle*, *meeting*, *outdoor*, *waterscape*, *flag* and—as in the case of the examples in Fig. 1—*people*. In the same way, videos about *court hearings* could be expected to include many indoor scenes in courtrooms, while videos about *zoos* could be expected to depict animals significantly more often than other visual concepts. Videos about *celebrations* and *politics* typically contain shots involving people, but with different occurrence patterns: frequent appearance of larger groups of people might be more typical in case of celebration, whereas a video about politics would include more shots of individual people (e.g., taken during interviews with individual politicians).

In view of the above, the information on visual concepts should not go unexploited for the purpose of retrieving videos based on semantic themes. While this information remains insufficient to link a video directly to a topical query, we foresee a large value of this information in its ability to help determine whether two videos are similar in terms of their semantic themes. As we also conjecture that the visual concept detectors have a potential to encode information about stylistic features related to television production rules [16], their value for determining video similarity may expand across a broad range of semantic themes defined at various abstraction levels.

We propose in this paper a retrieval approach that consists of the following two steps:

- Building a video representation that is suitable for assessing similarity between two videos in terms of their semantic themes and that is based on aggregating the outputs of visual concept detectors across different shots of a video.
- Query expansion selection (QES) that responds to topical queries and that is based on the query performance prediction (QPP) principle (e.g., [3,35]). Here, the proposed video representation serves as input into query performance indicators, which evaluate various results lists produced by different query modifications.

The list with the highest estimated performance is then adopted as the best possible search result given a topical query, video collection, available search mechanisms and the resources for query modification.

The main research questions we address in this paper are the following:

- To which extent can the proposed QES retrieval approach outperform a baseline system that solely relies on the spoken content channel?
- For which categories or abstraction levels of semantic themes does the QES approach work well and what reasons of failure can be inferred for semantic themes for which the approach fails?
- Is it possible to obtain a more reliable prediction through combining concept-based indicators and text-based indicators of query performance?

We first explain the rationale and outline the contribution of our retrieval approach in Sect. 2, while in Sect. 3, we provide an insight into the state-of-the-art in the main technologies underlying this approach. Then, we introduce the two main approach steps listed above, namely building the video representation that we refer to as *Concept Vector* (Sect. 4) and designing the QES retrieval framework utilizing this video representation (Sect. 5). Sections 6, 7 and 8 are dedicated to the experimental evaluation of our approach. Sections 6 and 7 address the first two research questions mentioned above, while the third research question is addressed in Sect. 8. The discussion in Sect. 9 concludes the paper.

2 Approach rationale and contribution

We base our approach on the same rationale that is underlying general QPP approaches [3,7,35] and which builds on the clustering theorem [20] stating that closely related documents tend to be relevant to the same request. In our approach, for analysing the relatedness between videos in terms of a semantic theme, we rely on the discussion in Sect. 1 and propose a video representation that exploits general distribution patterns of a large set of visual concepts detected in a video. Hereby, 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 the assurance that the concept set used provides complete semantic coverage of the visual content of the collection. The possibility to work with a general set of visual concept detectors makes our retrieval approach unsupervised, and therefore, opens a broader search range than in the case of supervised alternatives. Examples of such alternatives are the approaches that learn or otherwise generate mappings between specific visual concepts and semantic themes. Such approaches, which have been studied for shot-level retrieval, cf. [10,26], face the challenge of collecting a sufficiently large and representative set of visual concepts, particularly daunting for never-before-seen topical queries and being rather sensitive to the quality of visual concept detectors. Furthermore, as discussed in more detail in Sect. 3.3, such approaches are commonly tailored for TRECVID-like queries, differing from semantic themes in their strong reference to the visual channel of the video. In addition, since statistical information is collected over a large set of concept detectors, our approach is less sensitive to noise in the individual detectors.

Different results lists serving as input to query performance prediction are obtained for different query expansions created by adding additional terms to the original query. Query expansion (see Sect. 3.1 for more information) is widely deployed in the field of information retrieval (IR) to enrich the original query so as to provide a better match with documents in the target collection. In particular, it is known to increase recall [15]. In the area of spoken content retrieval, query expansion is often used [12,33] where it also compensates for errors in the speech recognition transcripts. The danger of query expansion is, however, that it may introduce inappropriate terms into the query, causing topical drift. Given an initial query text, a speech transcript of a video collection and a set of search results lists obtained for different query expansion methods and applied to the speech transcript, our QES approach controls the drift and selects the most appropriate query expansion method.

In our previous work [23], coherence indicators of query performance, exploiting pair-wise video similarities in terms of their spoken content, demonstrated the ability to improve retrieval at the semantic theme level within the proposed QES framework. In this paper, we revisit and adjust this framework to first investigate to which extent a modification of these *text-based* coherence indicators into the indicators exploiting *concept-based* similarities between videos can lead to an improvement of the semantic-theme-based video retrieval within the QES framework. Then, we also investigate whether additional improvement could be achieved by combining text-based and concept-based indicators.

In addition to being the first work to address in depth the problem of semantic-theme-based video retrieval, the main novel technical contribution of our approach is an integration of the output of visual-concept detectors aggregated across the entire video and the output of automatic speech recognition, both known to be noisy. We will show that through such integration, an overall improvement in retrieving videos using topical queries can be achieved, compared to several baseline approaches commonly used in the IR field. More specifically, we will demonstrate that for a given query, our concept-based query performance indicators are indeed effective in selecting the best out of available search results lists. Finally, we will show that a simple combination of concept-based indicators with the text-based alternatives might significantly improve performance in terms of mean average precision (MAP) and that, more importantly, a combined coherence indicator selects the optimal results list in over 35 % of queries, more than state-of-the-art text-based indicators.

3 Related work

3.1 Query expansion

A common problem in information retrieval is a mismatch between vocabularies of the query and the collection being queried. This problem is often addressed by expanding the query using, for instance, pseudo-relevance feedback or thesauri. Query expansion can be particularly helpful in the case of spoken content retrieval in which speech recognizer errors, and particularly, errors caused by words spoken in the recognizer input, but missing in the recognizer vocabulary, frequently occur. It is sometimes difficult to separate the improvement contributed by the expansion itself from the error compensating effects, but overall query expansion is known to yield improvement [12,33]. For example, recognizer error occurring for the original query term *excavation* might be compensated by expanding the query with additional related terms, such as *digging*, *archaeology*, *archaeologist* and *artifacts*, which are potentially correctly recognized. Although proper query expansion may generally improve the retrieval results, it also introduces the danger of a topical drift [14], the tendency of expanded query to move away from the topic expressed by the original query.

3.2 Query performance prediction

Topical drift can be controlled by appropriate query performance prediction applied to decide whether a query should be expanded and how [4]. 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. In [3], query prediction uses the Kullback–Leibler divergence between the query model and the background collection model (clarity score). Yom-Tov et al. [35] proposed efficient and robust methods for query performance prediction based on measuring the overlap between the results returned for a full query and its sub-queries.

Recently, a coherence-based approach to query prediction has been proposed [7]. This approach measures the topical coherence of top documents in the results list to predict query performance. The approach is low in computational complexity and requires no labeled training data. Further, the coherence-based approach is appealing, because it goes beyond measuring the similarity of the top documents in a results list to measuring their topical clustering structure [8]. The coherence score is thus able to identify a results list as high-quality even in the face of relatively large diversity among the topical clusters in the top of results list.

In our recent work [23], we demonstrated the performance of the coherence score defined in [8] and two light-weight alternatives for the task of text-based QES. Subsequently, we carried out initial work, reported briefly in [22,24], which established the potential of coherence score to be useful for multimodal QES. In this paper, we present the fully developed version of that initial approach including automatic generation of Concept Vectors for video representation, combining the proposed text-based and concept-based query performance indicators and validation on a large dataset.

In [31], an approach to performance comparison of web image search results has been proposed. The underlying ideas, including assumptions on density of relevant and non-relevant images and their pairwise similarities place this approach into the group of coherence-based approaches. However, it requires training and relies on preference learning, which could eventually reduce applicability to unseen queries. In addition, the set of queries used in the experiments indicates a strong reference to the visual channel and it remains unclear whether the approach could be applied for multimedia information retrieval at a higher semantic level, especially since the models were built based on low-level visual features only.

3.3 Multimodal video retrieval

Since a video conventionally consists of both a visual and audio track, multimodal approaches are clearly necessary to exploit all available information to benefit video retrieval. Our QES approach bears closest resemblance to reranking approaches, which use visual information to refine the initial results returned by spoken content retrieval [9,21,30]. However, there are important differences between QES and reranking. First, reranking approaches are restricted to reordering the initial results list—there is no mechanism that allows visual features to help admit additional results. Second, reranking methods are static and therefore known to benefit some queries and not others [9,21,30], while our QES approach adapts itself to queries. It attempts to maximally exploit the available information to select the best results list per query.

Another important difference between the work presented here and the previous work is the type of the retrieval task. As noted in the Sect. 1, semantic-theme-based video retrieval involves retrieving video according to its subject matter. Typical semantic theme (topical) queries are thus defined at a higher abstraction level and therefore substantially different from conventional TRECVID queries, which include named persons, named objects, general objects, scenes and sports (cf. [6]). TRECVID-type queries are strongly related to the visual channel and may not be actually representative of the overall topic of the video. 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., [9,18,28,30]), we consider entire videos as retrieval units. Our decision to move beyond shot-level retrieval is guided by the reasoning that a semantic theme is an attribute of either an entire video or a video segment of a significant length. We also believe that in many real-world search scenarios, e.g., popular content sharing websites, such as *YouTube* and *blip.tv*, users are actually looking for the entire videos to watch and that clips or segments must be of a certain minimum length to satisfy users' information need. While there has been little effort in the past that targeted video retrieval beyond the level of individual shots, recently, a story-level video retrieval approach was proposed that retrieves news items containing visually relevant shots [1]. Although relevance is not assessed with respect to the semantic theme, we mention this approach here because it is similar to our own regarding a relatively large retrieval unit and also the use of language models built over the concept detector output.

The increasing awareness of the need to address queries at a higher abstraction level than, e.g., LSCOM, can also be observed from the reformulation of a TRECVID search task, which was renamed to *known item search task* in TRECVID 2010 [19] and which included a limited number of theme-based queries, as well as a new video-level retrieval evaluation metric.

4 Building concept vectors

In this section, we present our approach for automatically creating Concept Vectors, visual concept-based representations of videos that are used to calculate similarities between videos that capture resemblances in terms of a semantic theme.

4.1 Making use of incomplete sets of noisy visual concept detectors

Since the relation between the semantic theme of a video and its visual content is potentially weak, the problem of successfully utilizing the visual channel for the purpose of query performance prediction appears to be rather challenging. In view of the discussion in Sect. 1, we believe that the intermediate representation at the level of visual concepts could lead to a solution for this task. Like words in the speech channel, concepts in the visual channel can be said to reflect the elements of the thematic content of the video.

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 [28]. Our approach is based on the insight that in spite of a relatively poor performance and noisiness of individual visual concept detectors at the shot level, aggregating the results of concept detections across a series of shots could help reduce the influence of this noise and still provide the basis for a reasonable video-level representation in the use context addressed in this paper.

The question has been raised in the literature of how many and which concept detectors would be required to sufficiently cover the entire semantic space for the purpose of effective video retrieval in a general use case [5]. Although, ideally, as many concept detectors as possible should be available to be able to handle enormous diversity of visual content and address a broad range of video search requests, the reality is that the set of available concept detectors will always be limited and not necessarily representative for every content domain. We hypothesize, however, that availability of the optimal visual concept set for a given use case is not critical for successful deployment of our approach, provided that mechanisms are developed to determine which particular concepts from the available concept set are more informative to be applied on a particular video collection.

Based on the above two hypotheses, we approach automatic generation of Concept Vectors by starting from an arbitrary set of available visual concept detectors, analyzing their output and selecting the most representative (informative and discriminative) visual concepts. Technical steps of this approach are described in more detailed in the subsequent parts of this section.

4.2 Concept-based video representation

To create our Concept Vectors, we follow the general process illustrated in Fig. 2 in which we draw an analogy to the conventional information retrieval and consider visual concepts as terms in a video “document”. In this process, we aim at representing a video v from a collection V using a vector $\mathbf{x}_v$ defined as

$$\mathbf{x}_v = [x_{1v}, x_{2v}, \ldots, x_{|C|v}]^\top \tag{1}$$

where x_{cv} is the weight of the concept c in video v, C is a general set of visual concepts and $^\top$ is the transpose operator. The weight x_{cv} serves to indicate the importance of the concept c in representing the video v. In the conventional information retrieval, this importance is generally expressed as a function of the $\mathrm{tf}_{c,v}$ (term frequency) and idf_c (inverse document frequency) [25], which reflects the number of occurrences of a term in a video and its discriminative power within the collection, respectively. The index “c, v” indicates that the TF component of the weight is specific for a video, while the index “c” reflects that the IDF component is computed over the entire collection.

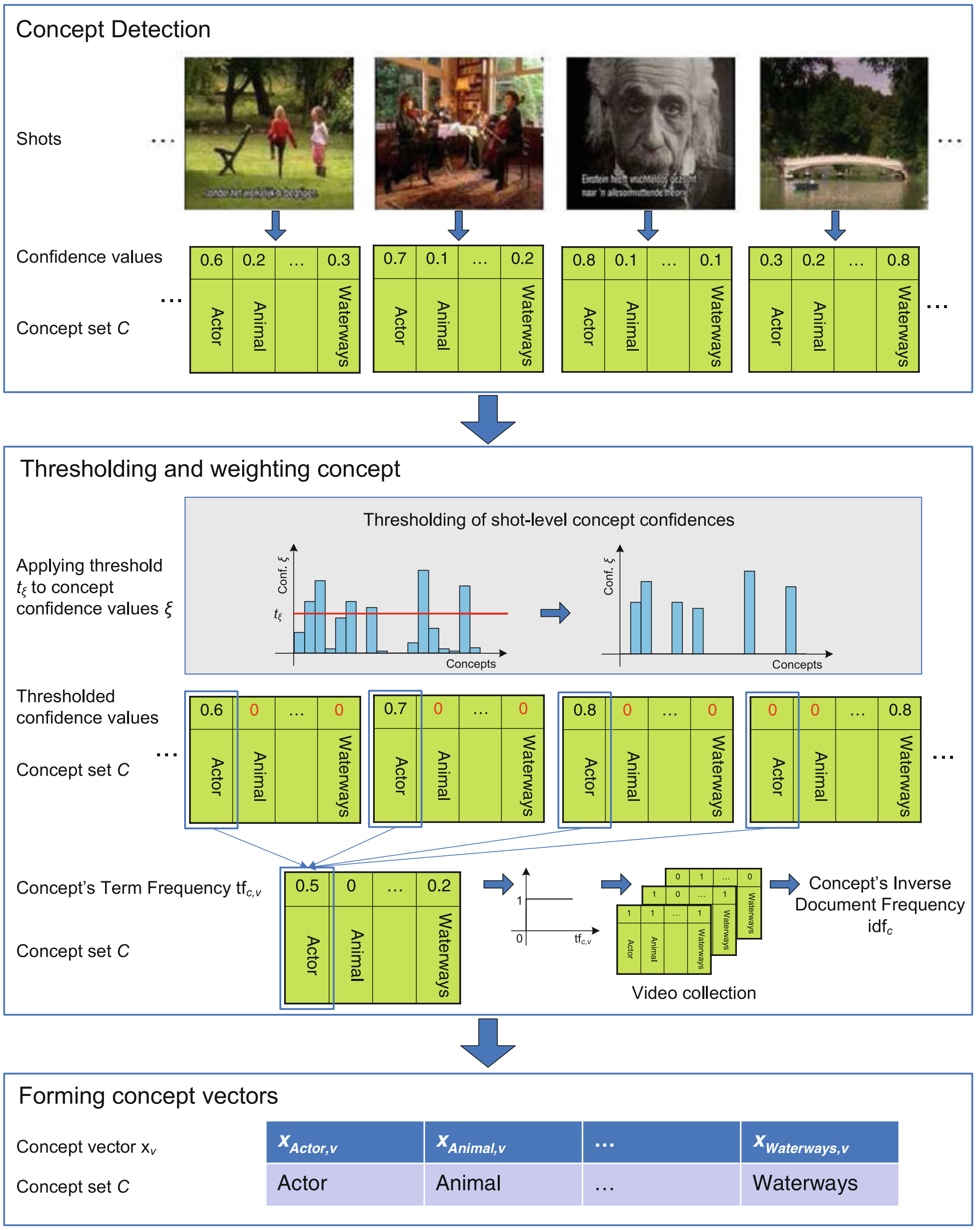
When computing the $\mathrm{tf}_{c,v}$ component of the weight, we take into account the fact that state-of-the-art concept detection systems [11,27] usually output shot-level lists of confidence scores for the given visual concepts, rather than binary judgments. For this reason, we model the term frequency here by the sum of a concept’s confidence scores taken from each shot of a video. To avoid bias toward videos containing more shots, we normalize the sum of confidence scores with the number of shots. Furthermore, recent works (i.e., [11,27]) revealed that the values of visual concept confidence vary widely within the interval of [0, 1], with low confidence values commonly indicating erroneous detection. Low confidence values effectively introduce a large amount of noise into the system, which will negatively bias the computation of idf_c . Therefore, we analyze the outputs of individual concept detectors and consider the reliable outputs only. In other words, we perform thresholding at the shot level to retain only those concepts in the representation that have substantial confidence scores. In our approach, thresholding is an essential step also because, as revealed by our exploratory experiments, reliable indicator of term (concept) frequency is critical for reliably selecting a subset of representative concepts.

Taking into account the above considerations, we compute $\mathrm{tf}_{c,v}$ according to the following expression:

$$\mathrm{tf}_{c,v} = \frac{\sum_{j=1}^{N_v} \{\xi_{c,v,j} : \xi_{c,v,j} > t_\xi\}}{N_v} \tag{2}$$

Here, $\mathrm{tf}_{c,v}$ is the normalized frequency of a concept c in video v, N_v is the number of shots in a video and $\xi_{c,v,j}$ is

Fig. 2 Illustration of our approach to concept-based video representation starting from a general concept set C. Final concept vectors $\tilde{\mathbf{x}}_v$ are created based on the subset $\widetilde{C}$ of selected concepts, as explained in Sect. 4.3



the confidence of the presence of a particular concept c in the shot j of video v as provided by the concept detector. The value of the threshold t_ξ we introduce for the purpose of denoising the output of the concept detectors is not critical if selected above a certain value. In our experiments, threshold values larger than 0.3 yielded insignificant difference in the performance.

$$\mathrm{idf}_c = \log \frac{|V|}{|\{v : \mathrm{tf}_{c,v} > 0\}|} \tag{3}$$

While $\mathrm{tf}_{c,v}$ represents the intensity of concept occurrence in a single video, idf_c (c.f. (3)) serves to incorporate the general pattern of visual concept occurrence within the entire collection. idf_c is computed by first dividing the entire number of videos in the collection with the number of videos in which the given concept is present and then by taking a logarithm of the quotient. Different ways of mapping $\mathrm{tf}_{c,v}$ and idf_c onto x_{cv} will be investigated experimentally in Sect. 7.

4.3 Concept selection

The goal of concept selection is to choose a subset of concepts from the available set C that are able to capture semantic similarities between videos. Concept selection can be seen as a feature selection problem known from the pattern recognition and information retrieval domain. Through years, many methods have been proposed to select features [29,34], many of which are supervised and require prior training. Our

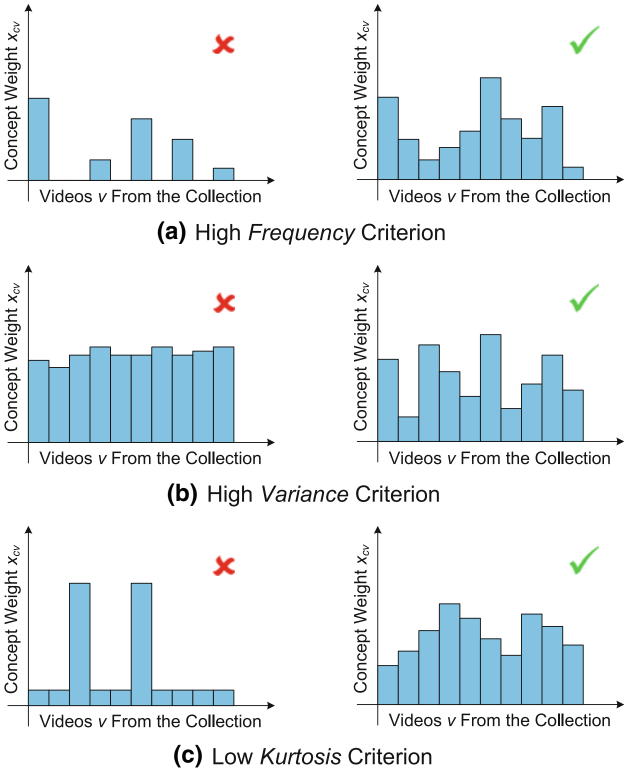


Fig. 3 Illustration of *frequency*, *variance* and *kurtosis* criteria for concept selection. Distribution examples on the right show the desired behavior of frequency, variance and kurtosis for marking relevant visual concepts

previous work revealed a high positive correlation between the frequency of concept occurrence across the collection and its effectiveness in discriminating between videos based on the semantic theme [24]. To have our approach completely data driven and unsupervised, we introduce a method for concept selection based on a simple heuristics that involves computing of the *frequency*, *variance* and *kurtosis* of visual concepts in the video collection. As will be explained in Sect. 6.2.2, here we set $x_{cv} = \mathrm{tf}_{c,v}$.

Frequency. We conjecture that concepts that occur in many videos within the collection will be more helpful in comparing videos than those concepts appearing in only few videos (Fig. 3a). Then, the relative difference in the importance weights of such concepts can provide a basis for calculating similarity between two videos. For each concept c we compute $freq_c$ by aggregating the concept counts a_{cv} across videos in which that concept appears:

$$freq_c = \sum_{v=1}^{|V|} a_{cv} \quad \text{and} \quad a_{cv} = \begin{cases} 1, & x_{cv} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Variance. Selecting the frequent concepts only is not enough, since some frequent concepts might have importance weights distributed uniformly throughout the collection. In that case, the concept will not be discriminative for comparing videos. Therefore, we require these frequent concepts to also have a high *variance* (Fig. 3b) of their importance weights across the video collection as well:

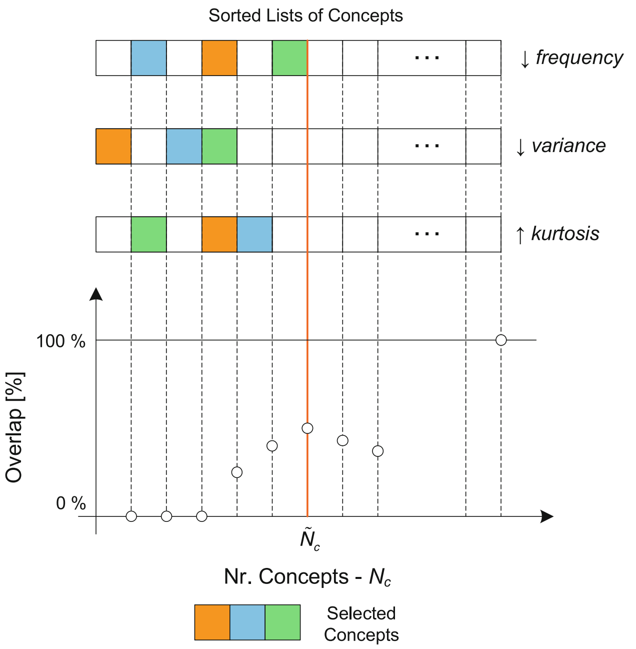


Fig. 4 Illustration of the procedure for selecting concepts that satisfy the frequency, variance and kurtosis criteria

$$var_c = \mathrm{var}(\mathbf{y}_c), \mathbf{y}_c = [x_{c1}, x_{c2}, \ldots, x_{c|V|}] \tag{5}$$

where $\mathbf{y}_c$ is the vector of weights of concept c in all videos in the collection.

Kurtosis. A high variance (5) might be the consequence of either infrequent extreme deviations or, preferably, frequent, but moderate variations of concept weights across the collection. To isolate the concepts with frequent but moderate variations, we focus on those concepts with a low kurtosis. Kurtosis is a measure of “peakedness” of the probability distribution of a real-valued random variable (Fig. 3c). We compute $kurt_c$ of a concept using (6), where μ and σ are the mean and the standard deviation of the vector $\mathbf{y}_c$:

$$kurt_c = \frac{\sum_{v=1}^{|V|} (x_{cv} - \mu)^4}{(|V| - 1)\sigma^4} \tag{6}$$

As illustrated in Fig. 4, we produce three ranked lists by sorting the concepts according to the decreasing frequency and variance and increasing kurtosis in the collection. Then, we compute the percentage of the overlap between the three top-N_c lists for the increasing number N_c of top-ranked concepts. The process stops at $\widetilde{N}_c$, when the first dominant local maximum in the overlap value curve is reached (e.g., overlap of more than 70 %), after which the concepts are

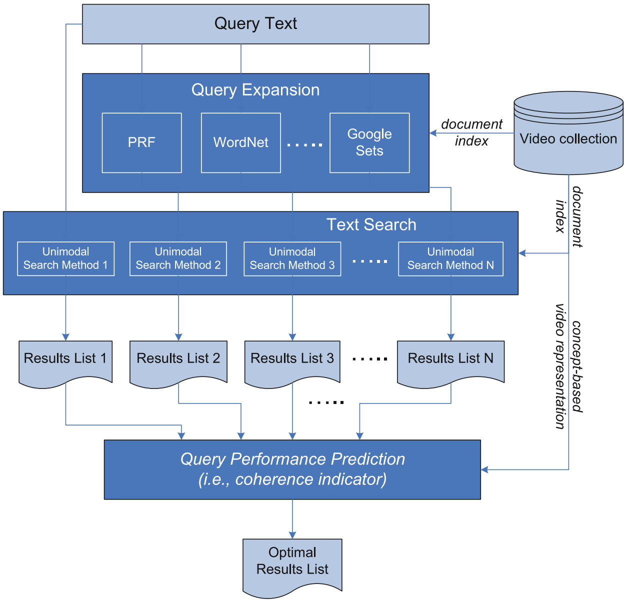


Fig. 5 Illustration of our QES approach

selected that are common to all three top-$\widetilde{N}_c$ lists. Prior to detecting local maxima, we smooth the overlap curve using a moving average filter, with the span parameter set to 10. The smoothing performed in this way helps reduce the influence of non-dominant local extrema and improves robustness of the concept selection approach. As will be shown in Sect. 6, the change in overlap with the increasing N_c remains largely consistent over different video collections and concept detection systems.

If we denote the three top-N_c lists of concepts sorted by *frequency*, *variance* and *kurtosis* as $Freq(N_c)$, $Var(N_c)$ and $Kurt(N_c)$, respectively, the selected set $\widetilde{C}$ of visual concepts can be defined as

$$\widetilde{C} = Freq(\widetilde{N}_c) \cap Var(\widetilde{N}_c) \cap Kurt(\widetilde{N}_c) \tag{7}$$

which leads to the "optimal" concept vector

$$\tilde{\mathbf{x}}_v = [\tilde{x}_{1v}, \tilde{x}_{2v}, \ldots, \tilde{x}_{|\widetilde{C}|v}]^\top \tag{8}$$

that serves as input for comparing videos in the subsequent query expansion selection step. In (8), $\tilde{x}_{cv}$ is the weight of concept $c \in \widetilde{C}$ in video $v \in V$ and $^\top$ is the transpose operator.

5 Query expansion selection

We approach the QES task from the data driven perspective, analyzing the collection being queried and the retrieval results list returned for the given query text. Figure 5 illustrates our QES approach. The system makes an unsupervised online analysis of the results lists produced for the original query and multiple query expansions to decide whether the query should be expanded, and if so, which of the alternative expansions would eventually yield the best results. An additional strength of our approach lies in the fact that we do not attempt to predict the retrieval performance (i.e., in terms of MAP) for each of the results lists (which usually requires prior training), but only compare the coherence of their top-ranked results. We evaluate three coherence indicators for this purpose, which will be introduced in the remainder of this section.

5.1 Coherence indicator

The coherence indicator [8] is used to select the results list with the highest coherence among the top-N retrieved results. The indicator is computed according to (9) as the ratio of video pairs in the top-N results whose similarity is larger than a threshold θ.

$$Co(TopN) = \frac{\sum_{u,v \in TopN, u \neq v} \delta(u,v)}{N(N-1)},$$
$$\delta(u,v) = \begin{cases} 1, & \text{sim}(\tilde{\mathbf{x}}_u, \tilde{\mathbf{x}}_v) > \theta \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

The threshold θ is set as a similarity value between particularly close videos in the collection. The threshold choice will be further discussed in Sects. 7 and 8. As a similarity measure $\text{sim}(\tilde{\mathbf{x}}_u, \tilde{\mathbf{x}}_v)$ we use the cosine similarity between the concept vectors (8) computed for videos u and v.

5.2 Max-AIS and mean-AIS indicators

The max-AIS and mean-AIS indicators [23] have been introduced as an alternative to the coherence score, because they do not need the reference to the video collection and make the decision based on the analysis of top-N ranked videos only. These indicators select the query expansion producing a results list in which top-N videos are characterized by high average item similarities (AIS) with their fellows. For video v AIS is computed according to (10).

$$AIS_v = \frac{\sum_{u \in TopN, u \neq v} \text{sim}\,(\tilde{\mathbf{x}}_u, \tilde{\mathbf{x}}_v)}{N-1} \tag{10}$$

Again, as a similarity measure $\text{sim}(\tilde{\mathbf{x}}_u, \tilde{\mathbf{x}}_v)$, we use the cosine similarity between the concept vectors (8) computed for videos u and v. Max-AIS indicator takes the maximum AIS value of all top-N videos in the results list, while mean-AIS takes the average of the AIS values in the top-N.

6 Experimental setup

This section describes our experimental framework and gives the implementation details of our approach.

6.1 Datasets

The experiments are performed on two datasets, that are reissues of the TRECVID 2007, 2008 and 2009 data made for the purposes of the "Tagging Task: Professional Version" offered for the MediaEval 2010[3] benchmark [13]. This benchmark also provided ground truth in the form of semantic theme labels assigned by professional archivists. The datasets are referred to as DS-1 and DS-2 and correspond to the MediaEval 2010 development and test dataset, respectively. Both datasets consist of the news magazine, science news, news reports, documentaries, educational programming and archival videos, provided by The Netherlands Institute for Sound and Vision (S&V)[4]. For the experiments, we use both DS-1 and DS-2 to investigate generalization of our approach across datasets. Unless stated otherwise, we do not treat them as the development and the test set, but rather as two equal datasets.

6.1.1 Description of DS-1 dataset

The DS-1 dataset is a large subset (nearly all) of the TRECVID 2007 and 2008 datasets, which consist of 219 videos each (438 videos in total). In the process of creating the DS-1 dataset, the videos without a semantic theme label were removed. Further, the videos without the automatic speech recognition transcripts and/or machine translation were also discarded. This led to a dataset consisting of 405 videos. As the queries, 37 semantic theme labels assigned by the S&V archive staff were used. These labels were selected such that each of them has more than five videos associated with it. The list of labels was post-processed by a normalization process that included standardization of the form of the labels and elimination of labels encoding the names of personages or video sources (e.g., amateur video).

6.1.2 Description of DS-2 dataset

The DS-2 dataset is composed of videos from TRECVID 2009 dataset. Only videos (400 in total) that did not occur in TRECVID 2007 and 2008 were included. Again, the videos without a semantic label provided by the S&V have been removed. Further, the videos without the automatic speech recognition transcripts and/or machine translation were also discarded. This led to a dataset consisting of 378 videos. As the queries, a set of 41 semantic labels assigned by the S&V archive staff were used, defined as explained in the previous section. As with the DS-1 dataset, the list of labels was post-processed by a normalization process that included standardization of the form of the label.

As shown in Tables 8 and 9, only 16 semantic labels are common to both DS-1 and DS-2 datasets, which serves to test the transferability of our approach to the never-before-seen queries. The performance stability across queries is analyzed in Sect. 7.6.

6.1.3 Query expansion methods

The query is modified using the following expansions.

- Conventional PRF (pseudo-relevance feedback), where 45 expansion terms are sampled from the automatic speech recognition transcripts of top-ranked videos in the initial results list produced for unexpanded query.
- WordNet expansion, by means of which the initial query terms are expanded with all their synonyms. The average total number of terms in such expanded queries is 12 for DS-1 and 13 for DS-2.
- Google Sets expansion, in which the initial query is expanded with a certain number of items (words or multi-word phrases) that frequently co-occur with that query on the web. To control topical drift, we limit the number of expansion items to 15.

6.2 Visual concept detectors

6.2.1 Concept detector choice

Videos from the DS-1 dataset are represented using CU-VIREO374 concept detection scores [11]. The system consists of 374 visual concepts selected from the LSCOM ontology [17]. To represent the DS-2 dataset, we used (separately) both CU-VIREO374 and MediaMill [27] visual concept detection scores for the purpose of comparative analysis. MediaMill system consists of 64 concept detectors and at the moment when the experiments described here were performed, their outputs were publicly available for DS-2 dataset only.

6.2.2 Concept selection procedure

We now experimentally verify the feasibility of the methodology for selecting a subset of representative visual concepts for a given collection, which is based on the frequency, variance and kurtosis of the concepts, as described in Sect. 4.3. In the experiments reported in Sect. 7, TF weighting yielded a better performance than TF–IDF in the concept selection task and therefore for the computation of concept frequency, variance and kurtosis, c.f. (4), (5) and (6), we set here $x_{cv} = \mathrm{tf}_{c,v}$.

Figure 6a–c shows the plots of frequency of concept occurrences, variance and kurtosis of $\mathrm{tf}_{c,v}$ throughout the video collection constructed using CU-VIREO374 concepts on the

[3] http://www.multimediaeval.org.

[4] http://www.beeldengeluid.nl.

DS-1 dataset and CU-VIREO374 and MediaMill concept detectors on the DS-2 dataset.

The results shown in Fig. 6a indicate that some concepts are present in almost all videos in the collection with a significant confidence, while a large subset of concepts appear only in a limited number of videos. This observation holds for both the DS-1 and the DS-2 dataset, and surprisingly, both for CU-VIREO374 and MediaMill concept detectors (not affected by the difference in number of concept detectors in both systems).

In addition, as shown in Fig. 6b, a small subset of concepts has a high variance in the DS-1/DS-2 dataset, while a larger number of concepts show relatively uniform values across the collection. Similar observation can also be made for kurtosis (Fig. 6c). The goal of our concept selection procedure is to isolate a set of concepts that appear as high as possible in the concept ranking (i.e., as far to the left as possible in Fig. 6a–c), meaning that they have high variance, high frequency and also low kurtosis. Finally, Fig. 6d shows the curves used to determine the length $\widetilde{N}_c$ of the ranked lists at which the set of selected concepts is generated as the overlap between the three lists. Supporting the illustration in Fig. 4, the curves indeed show clear local maxima at which $\widetilde{N}_c$ can be determined.

7 Experimental evaluation of QES based on concept-based coherence indicators

Through the experiments summarized in this section, we seek answers to the following research questions:

- How does the proposed QES approach perform if videos are represented using the original concept vector (1), without refinement through the concept selection step (Sect. 7.1)?
- To which extent does the concept selection step improve the QES performance and under which conditions (Sect. 7.2)?
- Do the results generalize onto a new dataset and under which conditions (Sect. 7.3)?
- What is the impact of the quality of visual concept detection on the QES performance (Sect. 7.4)?
- Is the performance gain stable across queries (Sect. 7.6)?

We will address these questions first by working with the coherence indicator (Co) defined in (9). The performance of alternative indicators (mean-AIS and max-AIS) is then analyzed separately in Sect. 7.5. Furthermore, we will evaluate our approach in view of the above questions through a comparative analysis involving the best-performing baseline approach. This reference approach is selected among the simple text-search baseline that uses speech recognition transcripts only and our three additional results lists produced using common query expansions (Sect. 6.1.3). The performance of these approaches in terms of MAP is shown in Table 1 for both DS-1 and DS-2 datasets. Note that the four results lists produced by these four baseline approaches are the ones that will be combined by our QES approach.

Table 1 MAP of the baseline and the query expansions used

Dataset	Baseline	PRF	WordNet	Google Sets
DS-1	0.2322	0.2619	0.1941	0.1271
DS-2	0.2381	0.2621	0.1867	0.1276

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 available baseline results lists. Instead, for each query, we target to always choose the best results list, whatever MAP it has. To make a comparison with the theoretical optimum, we make use of "oracle" indicators, the hypothetical indicators that always choose the correct query expansion. Here, we note that oracle indicators would achieve a MAP of 0.3082 and 0.3136 on the DS-1 and DS-2 datasets, respectively. These numbers can be seen as the upper limits of the achievable performance of our QES approach. Throughout experiments, we also compare the performance of our QES approach to the performance of the "Best Baseline", a baseline approach achieving the highest MAP for a given video collection.

Finally, the evaluation will take into account the influence of the main parameters of our approach:

- The threshold θ used for computing the Co indicator,
- top-N, the number of videos used for computing the Co, mean-AIS and max-AIS coherence indicators,
- the number of automatically selected visual concept detectors $|\widetilde{C}|$ and
- x_{cv}, that can be set to either TF or TF–IDF.

7.1 QES using all concepts

In the first query expansion selection (QES) experiment, we use both CU-VIREO374 and MediaMill concept detectors. In Table 2 we report the performance of the system for the optimal parameter settings.

Although the use of all visual concepts available improves results on DS-2 dataset, the improvement is achieved for only a limited number of parameter settings and therefore cannot be considered robust enough. Moreover, in the case of DS-1 dataset improvement is not obtained for any parameter setting and/or choice of indicator. This finding supports our hypothesis that the concept selection is a critical step in our approach.

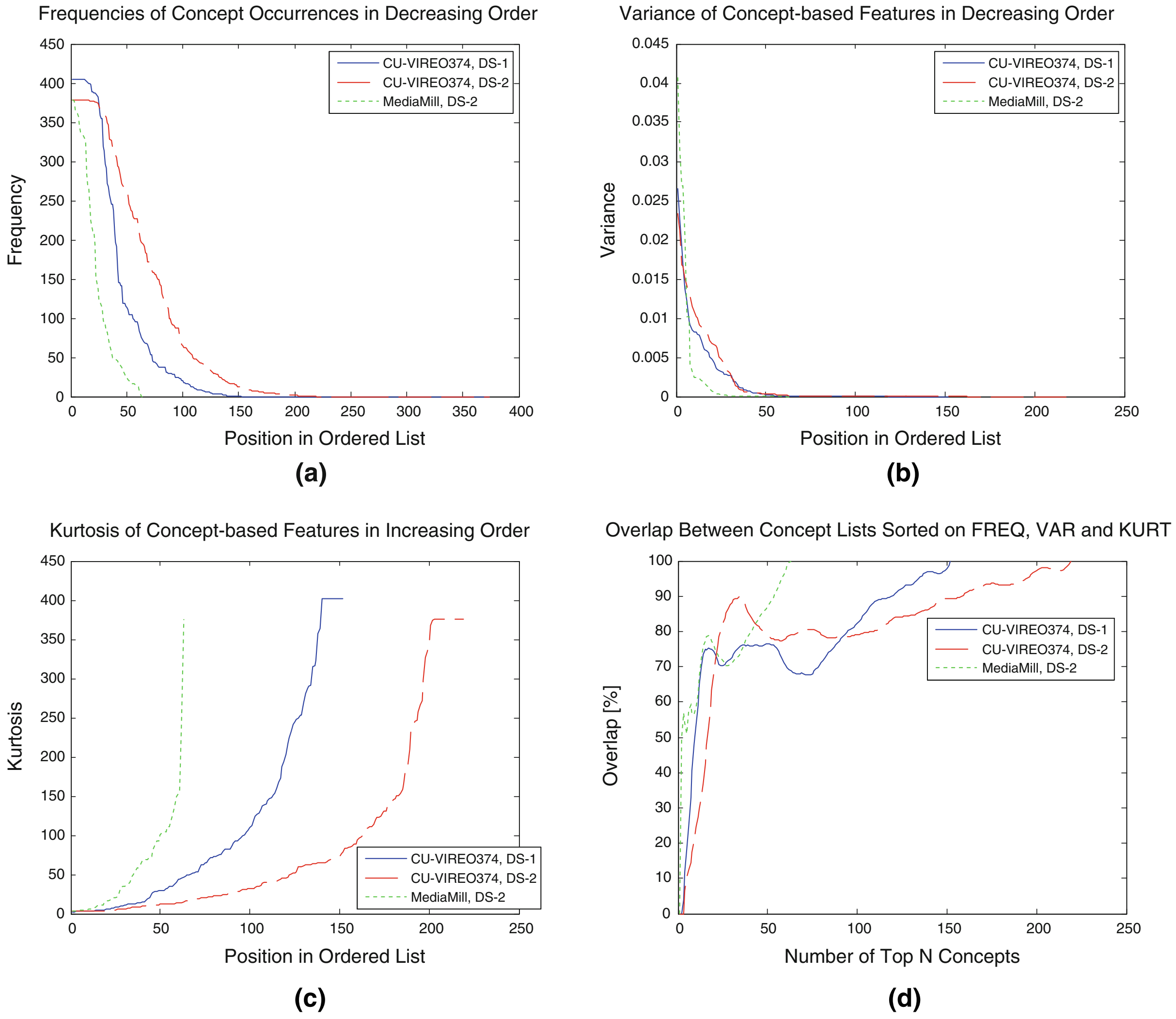


Fig. 6 **a** Frequencies of concept occurrences in the DS-1 and DS-2 datasets (sorted in decreasing order) **b** Concept variances (sorted in decreasing order) **c** Concept kurtoses (sorted in increasing order) **d** Percentage of overlap between the lists of concepts ordered according to frequency, variance and kurtosis

Table 2 MAP of our QES approach when all concepts and TF weights are used

Dataset	Concepts	Best Base.	QES	Oracle
DS-1	C-V374	0.2619	0.2363	0.3082
DS-2	C-V374	0.2621	0.268^	0.3136
DS-2	MM 64	0.2621	0.2743^	0.3136

Statistically significant improvement over the baseline is indicated with "^" (Wilcoxon Signed Rank test, $p = 0.05$)

7.2 QES applying the concept selection

In this section, we investigate the performance improvement that can be gained when applying concept selection. We first experiment with DS-1 and then analyze in Sect. 7.3 the capability of our approach to achieve a similar performance on the dataset DS-2 as well. For the DS-1 dataset from the entire CU-VIREO374 concept collection, only 15 most informative concepts are selected.

The performance of our QES approach in this case is summarized in Table 3. It is still far from the ideal performance of the oracle, but it shows a moderate improvement over the best-performing baseline, and also over the results in Table 2 (first row), where no concept selection was performed.

7.2.1 Robustness to parameter setting

To investigate the robustness of the retrieval performance in this case to parameter setting, we first investigate the QES performance for several values of threshold θ. In all cases, the number of top-N documents used to calculate the

Table 3 MAP of QES approach for DS-1 dataset when our concept selection approach on CU-VIREO 374 is used

Weights	Best Base.	QES	Oracle
TF	0.2619	0.2757^	0.3082
TF–IDF	0.2619	0.2648	0.3082

Statistically significant improvement over the baseline is indicated with "^" (Wilcoxon Signed Rank test, $p = 0.05$)

Table 4 MAP for different values of parameter θ

Weights	$\theta = 70$ %	$\theta = 80$ %	$\theta = 90$ %	Best Base.
TF	0.2735^	0.2745^	0.2757^	0.2619
TF–IDF	0.225	0.2648	0.23	0.2619

Statistically significant improvement over the baseline is indicated with "^" (Wilcoxon Signed Rank test, $p = 0.05$)

Table 5 MAP of QES approach for the DS-2 dataset when the concept selection approach on CU-VIREO374 concepts is used

Weights	Best Base.	QES	Oracle
TF	0.2621	0.2631	0.3136
TF–IDF	0.2621	0.256	0.3136

Statistically significant improvement over the baseline is indicated with "^" (Wilcoxon Signed Rank test, $p = 0.05$)

indicator is set to 20. This parameter value already yielded good results when calculating the coherence indicator on text vectors [23]. The results are shown in Table 4. Normalized TF video representation appears to be more robust to parameter setting than TF–IDF, since it shows consistent improvement for various values of parameter θ. In [8], the suggested parameter value is 95 %, but here it seems that the indicator calculated on concept-based features may be even more robust than the one calculated using conventional (text-based) TF or TF–IDF document representations.

Regarding the choice for x_{cv}, we investigate for which choice statistically significant improvements are obtained. In Tables 2, 3 and 4 the statistically significant improvements over the baseline retrieval method are indicated with "^". As a significance measure, we adopt the Wilcoxon signed rank test ($p = 0.05$), commonly used in information retrieval. As indicated in the tables, almost all improvements obtained when the TF weights are used are statistically significant, which supports our conclusion that they are indeed more valuable for our purposes. The superior performance of TF weights might be a result of the fact that it is the pattern of concept occurrence (reflected in TF) rather than the absolute presence or absence of a concept in videos (encoded by IDF) that provides more helpful means of capturing semantic similarity. This effect may be specific to the distribution of concepts within video, since in text retrieval, the IDF weight generally makes an important contribution. We conjecture that the IDF component is particularly sensitive to the noise of concept detectors and that a high IDF value for a particular concept might be caused by an erroneous detection. Finally, the reason for a lower performance might lay in the fact that we select a rather small subset of concepts that appear frequently in the collection, and thus the IDF component does not have a positive influence.

7.2.2 Optimality of the obtained results

Further, we analyze whether our concept selection approach is capable of selecting the optimal threshold for the number of concepts to be used. Here, we consider only TF weights, because, as shown in the previous section, they demonstrate a superior performance to TF–IDF weights. We gradually increase the number of top-N_c concepts in the lists produced based on frequency, variance and kurtosis criteria and thus the number of overlapping concepts. The best overall MAP of 0.2757 is obtained when 15 concepts are selected, which is the same result achieved with a concept set chosen using the automatically selected threshold. This finding confirms the capability of our approach to select the optimal value $\widetilde{N}_c$.

7.3 Generalization across datasets

For the DS-2 set, our concept selection approach extracts 32 representative concepts from the CU-VIREO374 concept collection. The performance comparison with the best-performing baseline approach and the oracle indicator is shown in Table 5. Measuring the performance of the system for the varying number of selected concepts, as described in the previous section, reveals that in the case of TF–IDF weights our approach indeed selects the optimal number of concepts. In the case of TF weights, the maximal performance (MAP = 0.27) is obtained using the coherence indicator on 15 selected concepts (similar to the DS-1 set).

Similarly to the DS-1 set, when TF representation is used, a moderate improvement is achieved. TF–IDF representation again appears to be less robust and here it performs even worse than the best baseline approach (still outperforming the other three baselines). When the selection approach is applied to MediaMill concept collection, a subset of 14 representative concepts is selected. As an illustration, the automatically selected concepts are: *Building*, *Crowd*, *Face*, *Hand*, *Outdoor*, *Person*, *PersonWalkingOrRunning*, *Road*, *Sky*, *Street*, *TwoPeople*, *Urban*, *Vegetation* and *Waterscape*. The performance of QES approach using the Co indicator is shown in Table 6.

Both TF and TF–IDF concept-based feature variants yield a modest performance improvement and again the TF

Table 6 MAP of QES approach for the DS-2 dataset when the concept selection approach on MediaMill concepts is used

Weights	Best Base.	QES	Oracle
TF	0.2621	0.2688^	0.3136
TF–IDF	0.2621	0.2673	0.3136

Statistically significant improvement over the baseline is indicated with "^" (Wilcoxon Signed Rank test, $p = 0.05$)

Table 7 MAP of QES approach for the DS-2 dataset when the concept selection approach on MediaMill concepts and the mean-AIS indicator are applied

Weights	Indicator	Best Base.	QES	Oracle
TF	mean-AIS	0.2621	0.2719^	0.3136
TF–IDF	mean-AIS	0.2621	0.27^	0.3136

Statistically significant improvement over the baseline is indicated with "^" (Wilcoxon Signed Rank test, $p = 0.05$)

weights are performing slightly better, which is consistent with our previous findings.

7.4 Impact of quality of concept detectors

Compared to CU-VIREO374, the use of MediaMill concept set yields an increased robustness to parameter settings and, as shown in Tables 2, 5, 6 and 7, generally gives a higher performance improvement (in terms of MAP). This is not unexpected, since the MediaMill system achieved the highest performance in TRECVID 2009 concept detection and interactive search tasks [28]. We can therefore conclude that the quality of concept detectors remains an important factor influencing the performance of our approach.

7.5 Alternative coherence indicators

In addition to the coherence indicator (Co), we test the usability of two alternative indicators of the topical clustering structure (mean-AIS and max-AIS). The experimental results show that when CU-VIREO374 concepts are used, the alternative indicators do not yield improvement on either DS-1 or the DS-2 set. However, when the MediaMill concepts are used in the DS-2 set, the overall best performer for a wide range of parameter settings is the mean-AIS indicator. Table 7 summarizes the performance of our QES system on the DS-2 set, when the automatic concept selection approach is applied to the MediaMill concept set.

Wilcoxon signed rank test reveals that the obtained improvements are indeed statistically significant. The results from Table 7 are also consistent with our earlier findings. Namely, in our previous work [23], we showed that the mean-AIS and max-AIS indicators might be successfully used for query expansion selection when the videos are represented by the vectors of TF–IDF weights calculated on the automatic speech recognition transcripts text of the videos only. Moreover, the overall best-performing indicator in those experiments appeared to be mean-AIS. We conjecture that the performance improvement can be attributed to a higher sensitivity of mean-AIS indicator to the quality of concept detectors.

Mean-AIS indicator calculated on TF and TF–IDF weights gives consistent improvement in performance for different sizes of top-N video set over which it is calculated (i.e., N = 5, 10, 15, 20) when MediaMill concepts are used. This finding confirms that, depending on the quality of concept detector set, our approach is relatively robust to parameter setting.

7.6 Performance stability across queries

In this section, we investigate how the improvement is distributed over queries. Our method predicts the correct query expansion in approximately 40 % of time, but it also seems to make the correct prediction in the critical cases where the AP improves significantly. The indicator seems to make the error generally only in the cases when the coherence of the tops of different results lists is similar, but fortunately, the MAP values of these lists are also similar. For example, after the failure analysis of the results presented in Table 7, when the mean-AIS indicator is used on TF weights, we concluded that our indicator chooses the correct expansion in 43.9 % of cases. Further analysis reveals that the indicator additionally chooses the second best expansion in 34.15 % of queries and the errors were generally made in the case where the second best and the best results lists have very similar coherence of top results. Basically, our indicator selects the best or second best expansion in roughly 78 % of queries. It is also important to note that our query expansion selection makes use of all available query expansions approaches. Namely, as shown in Table 1, the Google Sets expansion in general performs worse than the baseline retrieval, PRF and WordNet expansions, but there are queries for which it helps and our indicators seem to be capable of predicting such cases. For example, a failure analysis of the results presented in Table 6, when the Co indicator is used on TF weights, reveals that in the case of topical queries, such as *dictatorship* and *youth programs*, the Google Sets expansion yields the best results and our indicator appears to be capable of detecting it. Further, in the case of queries *youth programs* and *patients*, the best-performing expansions are Google Sets and WordNet, while the generally better-performing baseline and PRF expansion achieve 0 MAP. In both cases, our indicator manages to select the best query expansion.

As explained in the introduction, we expect the performance of our approach to be influenced by several important

Table 8 Successfulness of our concept-based indicator, text-based alternative indicator and the combined indicator in predicting the optimal query expansion on the DS-1 set

CU-VIREO374	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	+	-	+	+	-	+	-	+	+	-	-	+	+	-	+	-	+	-	-	+	-	+
Text	-	+	+	-	-	+	-	-	-	-	+	-	-	-	-	-	+	-	-	-	+	-	+	-	-	-	+	-	-	+	+	+	-	-	-	-	-
Combo	-	+	+	-	-	+	+	+	-	-	+	-	-	-	-	+	+	+	+	-	+	-	+	+	-	-	+	+	-	+	+	+	-	-	+	-	+
Semantic Labels	Work	Poverty	Biology	Foreign workers	Civil wars	Computers	Criminality	Cultural identity	Animals	Zoos	Economy	Ethnic minorities	Factories	Families	Celebrations	Fraud	Medicine	History	Brain	Assistance	Journalists	Children	Landscapes	Military personnel	Musicians	Music	Physics	Entrepreneurs	Wars	Seniors	Press	Politics	Court hearings	Elections	Food	Soccer	Scientific research

Table 9 Successfulness of our concept-based indicators, text-based alternative indicator and the combined indicator in predicting the optimal query expansion on the DS-2 set

MediaMill	+	-	-	-	-	-	+	+	-	+	-	-	-	+	-	-	+	+	-	+	+	-	-	-	-	-	+	-	-	+	-	+	-	-	-	-	+	+	-	+	-
CU-VIREO374	-	+	-	-	+	-	+	+	-	+	+	-	-	-	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	+	+	-	+	+	-	+	-	+	-	-	-
Text	-	+	+	-	-	-	+	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	-	+	-	-	-	-	+	-	+	+	+	-	-	+	-	-	-	+
Combo	-	+	+	-	-	-	+	+	-	+	-	-	-	+	-	-	+	-	-	+	+	-	-	-	-	-	+	-	-	+	-	+	+	+	-	-	+	+	-	+	+
Semantic Labels	Actors	Work	Asylum seekers	Biology	Books	Criminality	Daily life	Dictatorship	Animals	Economy	Ethnic minorities	Fiactories	Film	History	Health science	Brain	Youth	Youth programs	Judicial system	Children	Artists	Laboratories	Agriculture	Literature	People	Military personnel	Muslims	Wars	Seniors	Patients	Police	Politics	Paintings	Writers	Feature films	Refugees	Food	Women	Scientific research	Housing	Diseases

factors, such as abstraction level of a particular semantic label, semantic and visual diversity of the videos relevant to that semantic label and the quality of visual concept detectors used. Tables 8 and 9 show for which semantic labels (queries), our query performance prediction approach succeeds in selecting an optimal expansion. Here, for both datasets and concept sets, we use Co indicator on TF weights. A general observation can be made that the performance of our concept-based indicators is relatively independent of the abstraction level of a particular semantic label. In other words, the indicators manage to choose a correct results list for some more abstract (e.g., *daily life*, *politics*, *economy* and *history*) and some less abstract queries (e.g., *landscape* and *food*). We believe that in case of some abstract semantic themes, our concept-based indicators are able to capture high-level stylistic similarities between videos, originating in television production rules. For example, political documentaries and talk shows usually feature several people talking about the subject. Further, in Table 9, we observe that for some semantic labels MediaMill concept detectors perform better, while in some other cases, the better-performing concept detector set is CU-VIREO374. This may be attributed to the fact that many concepts selected for those sets are different and not all concepts are equally representative of a particular semantic theme. In addition, as shown in [11,27], performance of concept detectors varies significantly within a concept detector set, which further influences effectiveness and reliability of the set in capturing the semantic characteristics of a video. Finally, on the DS-2 set, our concept-based indicators perform well for some semantically related queries, such as *children*, *youth* and *youth programs*. This observation supports our assumption that the correct decisions of concept-based indicators actually do not occur randomly, but depend on the quality of concept detectors and the degree to which a particular semantic theme is visually constraining.

8 Comparison with the text-based and combined query performance indicators

While the experiments described in the previous section served to demonstrate that our concept-based video representation and the concept-based indicators of query performance

Table 10 MAP of QES approach for DS-1 and DS-2 set when the coherence indicator Co is used with concept-based and text-based video representations; performance of indicator selection method is shown as well

Dataset	Best Base.	QES concepts		QES text		QES concepts+Text		Oracle
		MAP	Corr. (%)	MAP	Corr. (%)	MAP	Corr. (%)	
DS-1	0.2619	0.2757^	40	0.2624^	30	0.2846^	54	0.3082^
DS-2	0.2621	0.2688^	37	0.2734^	32	0.2831^	44	0.3136^

Statistically significant improvement over the baseline is indicated with "^" (Wilcoxon Signed Rank test, $p = 0.05$)

are indeed promising solutions for semantic-theme-based video retrieval, here, we compare their performance with the performance of text-based alternatives. As discussed in Sect. 3, recently proposed coherence-based indicators (e.g., [7,8]), have been proven effective in a wide range of text information retrieval applications. In [23], we showed that post-retrieval coherence-based query performance indicators, such as those described in Sect. 5, might improve spoken content retrieval significantly. The retrieval framework used is similar to the one illustrated by Fig. 5, with, however, an important difference in video representation. For that, we exploit only the automatic speech recognition transcripts of the videos and represent each video as the vector of TF–IDF weights.

8.1 Text-based indicators on DS-1 and DS-2 datasets

In this experiment, we use DS-1 set for exploring the parameter space and report results on DS-2 set. To simplify the analysis of indicator fusion, we limit the experiments to the coherence indicator Co only and report cases in which the other indicators perform better in terms of MAP or robustness to parameter setting. We choose to focus on the coherence indicator in this experiment, also because, it is the only indicator to yield performance improvement on both DS-1 and DS-2 sets when CU-VIREO374 concepts are used to represent videos. For text-based video representation, we index English translation of the automatic speech recognition transcripts and create vectors of TF–IDF weights. Preprocessing includes stemming and rigorous stopword removal, where each word appearing in more than N_s% of videos is considered to be a stopword. Our exploratory experiments show that the best results are obtained for $N_s = 20$ %. Further, similarly to [8], we experimentally prove that the text-based coherence indicator yields optimal performance when computed on top-5 documents using a high value for document similarity threshold $\theta = 95$ %. The performance on both datasets is reported in Table 10.

The best performer on DS-1 set is our proposed max-AIS indicator, scoring an MAP of 0.2648 when computed on top-5 documents. In general, on both the DS-1 and DS-2 sets, max-AIS and mean-AIS appear to be more robust than the Co indicator, yielding improvement for a larger range of parameter settings. Finally, for the completeness of the analysis, we repeat the experiments representing videos as the vectors of normalized TF weights. Interestingly, normalized TF representation yields a similar performance improvement to TF–IDF for various values of stopword removal threshold $N_s \in [10\,\%, 90\,\%]$, which might suggest that some stylistic attributes of the conversational speech might be particularly useful for discriminating between videos based on the semantic theme.

8.2 Indicator selection

As discussed in Sect. 7.6, our best-performing concept-based indicator, mean-AIS, on DS-2 set chooses a correct query expansion in roughly 40 % of cases, while the first or second best expansion is selected in over 70 % of cases. Therefore, we expect that fusion of text-based and concept-based indicators might lead to a further performance improvement.

To prove the concept, we choose to perform fusion through a simple voting strategy, acknowledging that a more sophisticated fusion approach might yield a higher performance improvement. First, we compute the indicators for the results lists generated in response to the original query and the three query expansions used and then select to use a more confident indicator for that query. We consider an indicator as more confident if it has a larger relative difference δCo^m between outputs for the most coherent and second most coherent results list.

$$\delta Co^m = \frac{Co_1^m - Co_2^m}{Co_2^m}, m \in \{c, t\} \tag{11}$$

In (11), Co_1^c and Co_2^c are the outputs of the concept-based indicator computed for the most coherent and second most coherent results list, while Co_1^t and Co_2^t are the corresponding outputs of the text-based indicator.

Table 10 shows the performance of our QES approach when: (1) concept-based video representation (indicator) is used; (2) text-based video representation (indicator) is used; (3) a more confident out of two computed indicators is selected. In a separate field, for each indicator we show a percentage of correctly selected expansions (i.e., percentage of cases in which the optimal query expansion is selected).

As explained in the previous section, for the reasons of consistency and analysis simplification, we limit the experiment to coherence indicator Co only. On DS-1 dataset CU-VIREO374 concept set is used, while for DS-2 we make use of better-performing MediaMill visual concept detectors.

The results indicate that selection of a more confident indicator brings additional performance improvement in terms of both MAP and ratio of correctly selected query expansions, which proves our starting assumption. The results presented in Table 10 indicate that the concept-based indicators yield a comparable performance to the state-of-the-art alternatives in the IR field, computed using the spoken content only. Furthermore, in the case of concept-based indicators, performance improvement seems to be better distributed across queries (e.g., optimal results list/ query expansion is selected more often). An interesting observation can be made in Table 8: if either text-based or concept-based indicator manages to select the optimal results list, a combined indicator will succeed in the task as well. A similar, although not as constant, trend could be observed in Table 9, which further shows that even a simple combining of indicators can lead to a more reliable prediction. Finally, the experiments confirm our main assumption that the information relevant to a semantic theme can be extracted from the visual channel of the video and not only from its spoken content.

9 Discussion

We have presented an approach to semantic-theme-based video retrieval that uses shot-level outputs of visual concept detectors to automatically build video-level representations, here referred to as Concept Vectors. These vectors are used to calculate coherence indicators that enable query expansion selection (QES) within a post-retrieval query performance prediction framework. The novel contribution of our approach is the effective combination of the output of automatic speech recognition and visual-concept detection, both known to be noisy, to achieve an overall improvement in retrieval of videos according to the semantic theme specified by the query. Our approach does not aim at obtaining hypothetical maximum performance on the given datasets, but rather to select the best out of available results lists for a given topical query in an unsupervised fashion. Concept Vectors are used to compare videos with each other instead of with the query. In this way, we are able to avoid any training that would be necessary to create a step that maps the query onto the appropriate concepts. Therefore, our approach has a potential to be used in a larger number of applications than the alternative solutions based on e.g., supervised learning.

A key advantage of our approach is its ability to make effective use of the noisy output of concept detectors. In fact, our Concept Vectors are designed to make optimal use of a given set of concepts, meaning that we do not necessarily need a guarantee that the set of concepts that we use provides a complete coverage of the semantic space of the collection. However, the starting concept set should provide a certain minimum required semantic coverage necessary for discriminating between videos at the level of a semantic theme. In addition, given the concept detector sets of the same quality, the one providing a better semantic coverage is intuitively expected to yield a similar or better performance within our system.

Our experimental evaluation validated the effectiveness of our approach and confirmed that the automatic selection of concepts during the process of building the Concept Vector is critical for the retrieval performance improvement. Experiments also revealed that the automatically determined cut-off for the list of concepts to be used succeeds in approximating the optimal value. Further, it was shown that including the IDF factor provided no further performance gains, consistent with the conclusion that it is not so much the uniqueness of a concept in a video, but rather the frequency of that concept's appearance that best captures pair-wise similarity between videos in terms of semantic theme. The method for automatic selection of concepts to be used to build the Concept Vector was shown to be transferable in an unproblematic manner to an unseen dataset of a similar type. Changing datasets does, however, require a re-optimization of the parameters involved in calculating the coherence indicator, namely the θ cutoff and also the number of top-N documents used.

The improvement yielded by the approach is distributed relatively well across the board, i.e., its benefit is not localized to only certain types of queries. In particular, there is no apparent correlation between the absolute number of documents relevant to a particular query within the collection and the effectiveness of our QES approach. This observation supports our claim that the applicability of our approach generalizes well across different kinds of queries presented to the system, and in particular to new queries with new properties. The efficacy of the approach was shown to have a sensitivity to the quality of the concept detectors, with better-performing concept detectors yielding higher improvement of QES.

The automatic approach for generating Concept Vectors involves a relatively small number of concepts. If a small set of well-chosen concept detectors in certain scenarios, such as the one described in this paper, is sufficient to improve the results of semantic-theme-based retrieval, a productive avenue for the development of concept detectors is to concentrate on achieving high quality for a small number of detectors and not on training concept detectors that will cover the entire conceivable semantic space.

We demonstrate that not only spoken content of the video but also information extracted from the visual channel can be successfully exploited for discriminating between videos based on the semantic theme. Finally, here we show that a simple combination of "unimodal" coherence indicators of query performance, exploiting text-based and concept-based video representations, might further improve retrieval performance. More specifically, for each query, we first compute text-based and concept-based query performance indicators, and then, automatically select the more confident indicator to obtain a higher performance improvement, both in terms of overall MAP and percentage of correctly selected expansions. Experiments reveal that our combined query performance indicator makes a correct decision for 30 % queries more than a state of the art text-based alternative.

Our future work will involve investigation into the further refinement of the approach to building concept-based video-level representations. In particular, we are interested in exploiting not only the frequency of occurrence of concepts but rather 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 retrieval task.

References

1. Aly R, Doherty A, Hiemstra D, Smeaton A (2010) Beyond shot retrieval: searching for broadcast news items using language models of concepts. In: Advances in information retrieval, LNCS, vol 5993, Springer, Heidelberg, pp 241–252
2. Arijon D (1976) Grammar of the Film Language. Silman-James Press, Los Angeles, CA, USA
3. Cronen-Townsend S, Zhou Y, Croft WB (2002) Predicting query performance. In: Proceedings 25th annual international ACM SIGIR conference on research and development in information retrieval, ACM, SIGIR '02, pp 299–306
4. Hauff C, Murdock V, Baeza-Yates R (2008) Improved query difficulty prediction for the web. In: Proceedings 17th ACM conference on information and knowledge management, ACM, CIKM '08, pp 439–448
5. Hauptmann A, Yan R, Lin WH (2007) How many high-level concepts will fill the semantic gap in news video retrieval? In: Proceedings 6th ACM international conference on Image and video retrieval, CIVR '07, pp 627–634
6. Hauptmann A, Christel M, Yan R (2008) Video retrieval based on semantic concepts. In: Proceedings of the IEEE, 96(4):602–622
7. He J, Larson M, de Rijke M (2008) Using coherence-based measures to predict query difficulty. In: Advances in information retrieval, LNCS, vol 4956, Springer, Heidelberg, pp 689–694
8. He J, Weerkamp W, Larson M, de Rijke M (2009) An effective coherence measure to determine topical consistency in user-generated content. Int J Doc Anal Recognit 12:185–203
9. Hsu WH, Kennedy LS, Chang SF (2006) Video search reranking via information bottleneck principle. In: Proceedings 14th annual ACM international conference on Multimedia, ACM, MM '06, pp 35–44
10. Huurnink B, Hofmann K, de Rijke M (2008) Assessing concept selection for video retrieval. In: Proceedings 1st ACM international conference on Multimedia information retrieval, ACM, MIR '08, pp 459–466
11. Jiang YG, Yanagawa A, Chang SF, Ngo CW (2008) CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection. ADVENT Technical Report #223-2008-1, Columbia University, New York
12. Johnson SE, Jourlin P, Jones KS, Woodland PC (2000) Spoken document retrieval for trec-8 at cambridge university. In: Proceedings TREC-8, pp 197–206
13. Larson M, Soleymani M, Serdyukov P, Rudinac S, Wartena C, Murdock V, Friedland G, Ordelman R, Jones GJF (2011) Automatic tagging and geotagging in video collections and communities. In: Proceedings 1st ACM Internatinal Conference on Multimedia Retrieval, ACM, ICMR '11, pp 51:1–51:8
14. Lee KS, Croft WB, Allan J (2008) A cluster-based resampling method for pseudo-relevance feedback. In: Proceedings 31st annual international ACM SIGIR conference on research and development in information retrieval, ACM, SIGIR '08, pp 235–242
15. Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
16. Nack F, Dorai C, Venkatesh S (2001) Computational media aesthetics: finding meaning beautiful. Multimedia, IEEE 8(4):10–12
17. Naphade M, Smith JR, Tesic J, Chang SF, Hsu W, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE MultiMedia 13:86–91
18. Natsev AP, Haubold A, Tešić J, Xie L, Yan R (2007) Semantic concept-based query expansion and re-ranking for multimedia retrieval. In: Proceedings 15th international conference on Multimedia, ACM, MM '07, pp 991–1000
19. Over P, Awad G, Fiscus J, Antonishek B, Qu G (2010) TRECVID 2010 an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings TRECVID Workshop, NIST, pp 1–34
20. van Rijsbergen CJ (1979) Information retrieval. Butterworth.
21. Rudinac S, Larson M, Hanjalic A (2009) Exploiting visual reranking to improve pseudo-relevance feedback for spoken-content-based video retrieval. In: 10th Workshop on image analysis for multimedia interactive services, WIAMIS '09, pp 17–20
22. Rudinac S, Larson M, Hanjalic A (2010a) Exploiting noisy visual concept detection to improve spoken content based video retrieval. In: Proceedings ACM internatinal conference on Multimedia, ACM, MM '10, pp 727–730
23. Rudinac S, Larson M, Hanjalic A (2010b) Exploiting result consistency to select query expansions for spoken content retrieval. In: Advances in information retrieval, LNCS, vol 5993, Springer, Heidelberg, pp 645–648
24. Rudinac S, Larson M, Hanjalic A (2010c) Visual concept-based selection of query expansions for spoken content retrieval. In: Proceedings 33rd international ACM SIGIR conference on research and development in information retrieval, ACM, SIGIR '10, pp 891–892
25. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manage 24:513–523
26. Snoek CGM, Worring M (2009) Concept-based video retrieval. Founda Trends Inf Retr 4(2):215–322
27. Snoek CGM et al (2008) The MediaMill TRECVID 2008 semantic video search engine. In: Proceedings TRECVID Workshop
28. Snoek CGM et al (2009) The MediaMill TRECVID 2009 semantic video search engine. In: Proceedings TRECVID Workshop
29. Theodoridis S, Koutroumbas K (2008) Pattern Recognition, 4th edn. Academic Press, Waltham, MA, USA
30. Tian X, Yang L, Wang J, Yang Y, Wu X, Hua XS (2008) Bayesian video search reranking. In: Proceedings 16th ACM international conference on Multimedia, ACM, MM '08, pp 131–140

31. Tian X, Lu Y, Yang L, Tian Q (2011) Learning to judge image search results. In: Proceedings 19th ACM international conference on Multimedia, ACM, MM '11, pp 363–372
32. Vasconcelos N, Lippman A (1997) Towards semantically meaningful feature spaces for the characterization of video content. In: Proceedings international conference on image processing, IEEE Computer Society, ICIP '97
33. Woodland PC, Johnson SE, Jourlin P, Jones KS (2000) Effects of out of vocabulary words in spoken document retrieval. In: Proceedings 23rd annual international ACM SIGIR conference on research and development in information retrieval, ACM, SIGIR '00, pp 372–374
34. Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: Proceedings 14th international conference on machine learning, Morgan Kaufmann Publishers Inc., ICML '97, pp 412–420
35. Yom-Tov E, Fine S, Carmel D, Darlow A (2005) Learning to estimate query difficulty: including applications to missing content detection and distributed information retrival. In: Proceedings 28th annual international ACM SIGIR conference on research and development in information retrival, ACM, SIGIR'05, pp 512–519