Learned Lexicon-Driven Interactive Video Retrieval

Cees Snoek*, Marcel Worring, Dennis Koelma, and Arnold Smeulders

Intelligent Systems Lab Amsterdam, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{cgmsnoek, worring, koelma, smeulders}@science.uva.nl
http://www.mediamill.nl

Abstract. We combine in this paper automatic learning of a large lexicon of semantic concepts with traditional video retrieval methods into a novel approach to narrow the semantic gap. The core of the proposed solution is formed by the automatic detection of an unprecedented lexicon of 101 concepts. From there, we explore the combination of query-by-concept, query-by-example, query-by-keyword, and user interaction into the *MediaMill* semantic video search engine. We evaluate the search engine against the 2005 NIST TRECVID video retrieval benchmark, using an international broadcast news archive of 85 hours. Top ranking results show that the lexicon-driven search engine is highly effective for interactive video retrieval.

1 Introduction

For text collections, search technology has evolved to a mature level. The success has whet the appetite for retrieval from video repositories, yielding a proliferation of commercial video search engines. These systems often rely on filename and accompanying textual sources only. This approach is fruitful when a meticulous and complete description of the content is available. It ignores, however, the treasure of information available in the visual information stream. In contrast, the image retrieval research community has emphasized a visual-only analysis. It has resulted in a wide variety of efficient image and video retrieval systems e.g. [1,2,3]. A common denominator in these prototypes is their dependence on color, texture, shape, and spatiotemporal features for representing video. Users query an archive with stored features by employing visual examples. Based on user-interaction the query process is repeated until results are satisfactory. The visual query-by-example paradigm is an alternative for the textual query-by-keyword paradigm.

Unfortunately, techniques for image retrieval are not that effective yet in mining the semantics hidden in video archives. The main problem is the semantic gap between image representation and their interpretation by humans [4]. Where users seek high-level semantics, video search engine technology offers low-level abstractions of the data instead. In a quest to narrow the semantic gap, recent research efforts have concentrated on automatic detection of semantic concepts in video [5, 6, 7, 8]. Query-by-concept offers users an additional entrance to video archives.

* This research is sponsored by the BSIK MultimediaN project.

H. Sundaram et al. (Eds.): CIVR 2006, LNCS 4071, pp. 11–20, 2006.

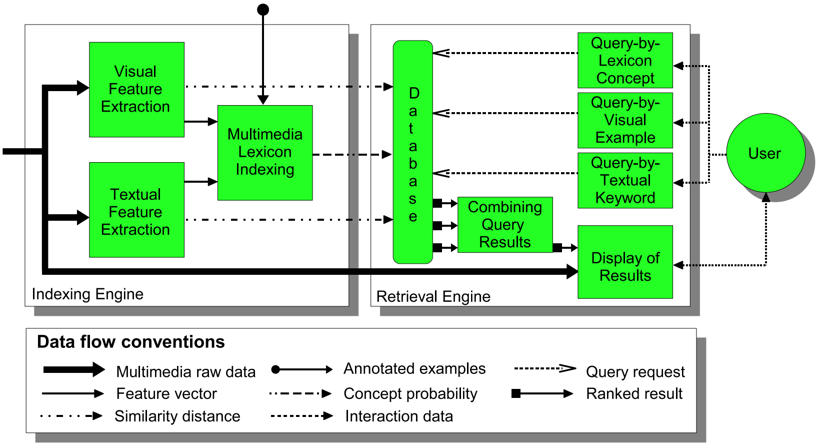

Fig. 1. General framework for an interactive video search engine. In the indexing engine, the system learns to detect a lexicon of semantic concepts. In addition, it computes similarity distances. A retrieval engine then allows for several query interfaces. The system combines requests and displays results to a user. Based on interaction a user refines search results until satisfaction.

State-of-the-art video search systems, e.g. [9,10,11,6], combine several query interfaces. Moreover, they are structured in a similar fashion. First, they include an engine that indexes video data on a visual, textual, and semantic level. Systems typically apply similarity functions to index the data in the visual and textual modality. Video search engines often employ a semantic indexing component to learn a lexicon of concepts, such as *outdoor*, *car*, and *sporting event*, and accompanying probability from provided examples. All indexes are typically stored in a database at the granularity of a video shot. A second component that all systems have in common is a retrieval engine, which offers users an access to the stored indexes and the video data. Key components here are an interface to select queries, e.g. query-by-keyword, query-by-example, and query-by concept, and the display of retrieved results. The retrieval engine handles the query requests, combines the results, and displays them to an interacting user. We visualize a general framework for interactive video search engines in Fig. 1.

While proposed solutions for effective video search engines share similar components, they stress different elements in reaching their goal. Rautiainen *et al.* [9] present an approach that emphasizes combination of query results. They extend query-by-keyword on speech transcripts with query-by-example. In addition, they explore how a limited lexicon of 15 learned concepts may contribute to retrieval results. As the authors indicate, inclusion of more accurate concept detectors would improve retrieval results. The web-based MARVEL system extends classical query possibilities with an automatically indexed lexicon of 17 semantic concepts, facilitating query-by-concept with good accuracy [6]. In spite of this lexicon, however, interactive retrieval results are not competitive with [10,11]. This indicates that much is to be gained when, in addition to query-by-concept, query-by-keyword, and query-by-example, more advanced interfaces for query selection and display of results are exploited for interaction.

Christel *et al.* [10] explain their success in interactive video retrieval as a consequence of using storyboards, i.e. a grid of key frame results that are related to a keyword-based query. Adcock *et al.* [11] also argue that search results should be presented in semantically meaningful units. They stress this by presenting query results as story key frame collages in the user interface. We adopt, extend, and generalize the above solutions.

The availability of gradually increasing concept lexicons, of varying quality, raises the question: how to take advantage of query-by-concept for effective interactive video retrieval? We advocate that the ideal video search engine should emphasize off-line learning of a large lexicon of concepts, based on automatic multimedia analysis, for the initial search. Then, the ideal system should employ query-by-example, query-by-keyword, and interaction with an advanced user interface to refine the search until satisfaction. To that end, we propose the *MediaMill* semantic video search engine. The uniqueness of the proposed system lies in its emphasis on automatic learning of a lexicon of concepts. When the indexed lexicon is exploited for query-by-concept and combined with query-by-keyword, query-by-example, and interactive filtering using an advanced user interface, a powerful video search engine emerges. To demonstrate the effectiveness of our approach, the interactive search experiments are evaluated within the 2005 NIST TRECVID video retrieval benchmark [12].

The organization of this paper is as follows. First, we present our semantic video search engine in Section 2. We describe the experimental setup in which we evaluated our search engine in Section 3. We present results in Section 4.

2 The MediaMill Semantic Video Search Engine

We propose a lexicon-driven video search engine to equip users with semantic access to video archives. The aim is to retrieve from a video archive, composed of n unique shots, the best possible answer set in response to a user information need. To that end, the search engine combines learning of a large lexicon with query-by-keyword, query-by-example, and interaction. The system architecture of the search engine follows the general framework as sketched in Fig. 1. We now explain the various components of the search engine in more detail.

2.1 Indexing Engine

Multimedia Lexicon Indexing. Generic semantic video indexing is required to obtain a large concept lexicon. In literature, several approaches are proposed [5, 6, 7, 8]. The utility of supervised learning in combination with multimedia content analysis has proven to be successful, with recent extensions to include video production style [7] and the insight that concepts often co-occur in context [5, 6]. We combine these successful approaches into an integrated video indexing architecture, exploiting the idea that the essence of produced video is its creation by an author. Style is used to stress the semantics of the message, and to guide the audience in its interpretation. In the end, video aims at an effective semantic communication. All of this taken together, the main focus of generic semantic indexing must be to reverse this authoring process, for which we proposed the semantic pathfinder [7].

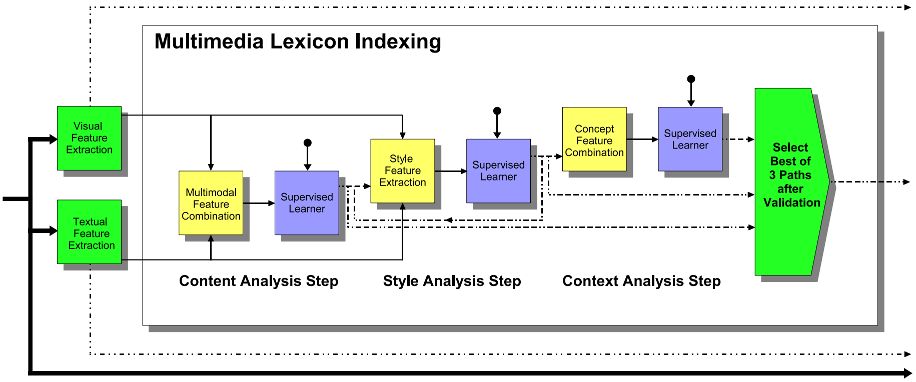
Multimedia Lexicon Indexing
Visual Feature Extraction
Textual Feature Extraction
Multimodal Feature Combination
Supervised Learner
Style Feature Extraction
Supervised Learner
Concept Feature Combination
Supervised Learner
Select Best of 3 Paths after Validation
Content Analysis Step
Style Analysis Step
Context Analysis Step

Fig. 2. Multimedia lexicon indexing is based on the semantic pathfinder [7]. In the detail from Fig. 1 we highlight its successive analysis steps. The semantic pathfinder selects for each concept a best path after validation.

The semantic pathfinder is composed of three analysis steps, see Fig. 2. The output of an analysis step in the pathfinder forms the input for the next one. We build this architecture on machine learning of concepts for the robust detection of semantics. The semantic pathfinder starts in the *content analysis step*. In this stage, it follows a data-driven approach of indexing semantics. It analyzes both the visual data and textual data to extract features. In the learning phase, it applies a support vector machine to learn concept probabilities. The *style analysis step* addresses the elements of video production, related to the style of the author, by several style-related detectors, i.e. related to layout, content, capture, and context. They include shot length, frequent speakers, camera distance, faces, and motion. At their core, these detectors are based on visual and textual features also. Again, a support vector machine classifier is applied to learn style probabilities. Finally, in the *context analysis step*, the probabilities obtained in the style analysis step are fused into a context vector. Then, again a support vector machine classifier is applied to learn concepts. Some concepts, like *vegetation*, have their emphasis on content thus style and context do not add much. In contrast, more complex events, like *people walking*, profit from incremental adaptation of the analysis by using concepts like *athletic game* in their context. The semantic pathfinder allows for generic video indexing by automatically selecting the best path of analysis steps on a per-concept basis.

Textual and Visual Feature Extraction. To arrive at a similarity distance for the textual modality we first derive words from automatic speech recognition results. We remove common stop words using the SMART's English stop list [13]. We then construct a high dimensional vector space based on all remaining transcribed words. We rely on latent semantic indexing [14] to reduce the search space to 400 dimensions. While doing so, the method takes co-occurrence of related words into account by projecting them onto the same dimension. The rationale is that this reduced space is a better representation of the search space. When users exploit query-by-keyword as similarity measure, the terms of the query are placed in the same reduced dimensional space. The most similar shots, viz. the ones closest to the query in that space, are returned, regardless of whether they contain the original query terms. In the visual modality the similarity query is by example. For all key frames in the video archive, we compute the perceptually uniform *Lab* color histogram using 32 bins for each color channel. Users compare key frames with Euclidean histogram distance.

2.2 Retrieval Engine

To shield the user from technical complexity, while at the same time offering increased efficiency, we store all computed indexes in a database. Users interact with the search engine based on query interfaces. Each query interface acts as a ranking operator on the multimedia archive. After a user issues a query it is processed and combined into a final result, which is presented to the user.

Query Selection. The set of concepts in the lexicon forms the basis for interactive selection of query results. Users may rely on direct query-by-concept for search topics related to concepts from this lexicon. This is an enormous advantage for the precision of the search. Users can also make a first selection when a query includes a super-class or a sub-class of a concept in the lexicon. For example, when searching for *sports* one can use the available concepts *tennis*, *soccer*, *baseball*, and *golf* from a lexicon. In a similar fashion, users may exploit a query on *animal* to retrieve footage related to *ice bear*. For search topics not covered by the concepts in the lexicon, users have to rely on query-by-keyword and query-by-example. Applying query-by-keyword in isolation allows users to find very specific topics if they are mentioned in the transcription from automatic speech recognition. Based on query-by-example, on either provided or retrieved image frames, key frames that exhibit a similar color distribution can augment results further. This is especially fruitful for repetitive key frames that contain similar visual content throughout the archive, such as previews, graphics, and commercials. Naturally, the search engine offers users the possibility to combine query interfaces. This is helpful when a concept is too general and needs refinement. For example when searching for Microsoft stock quotes, a user may combine query-by-concept *stock quotes* with query-by-keyword *Microsoft*. While doing so, the search engine exploits both the semantic indexes and the textual and visual similarity distances.

Combining Query Results. To rank results, query-by-concept exploits semantic probabilities, while query-by-keyword and query-by-example use similarity distances. When users mix query interfaces, and hence several numerical scores, this introduces the question how to combine the results. In [10], query-by-concept is applied after query-by-keyword. The disadvantage of this approach is the dependence on keywords for initial search. Because the visual content is often not reflected in the associated text, user-interaction with this restricted answer set results in limited semantic access. Hence, we opt for a combination method exploiting query results in parallel. Rankings offer us a comparable output across various query results. Therefore, we employ a standard approach using linear rank normalization [15] to combine query results.

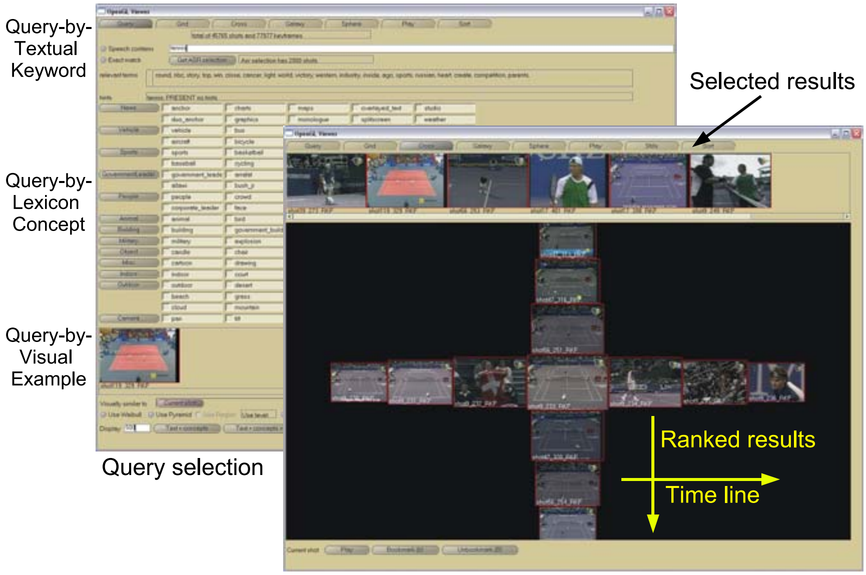

Fig. 3. Interface of the *MediaMill* semantic video search engine. The system allows for interactive query-by-concept using a large lexicon. In addition, it facilitates query-by-keyword, and query-by-example. Results are presented in a cross browser.

Display of Results. Ranking is a linear ordering, so ideally should be visualized as such. This leaves room to use the other dimension for visualization of the chronological series, or story, of the video program from which a key frame selected. This makes sense as frequently other items in the same broadcast are relevant to a query also [10, 11]. The resulting *cross browser* facilitates quick selection of relevant results. If requested, playback of specific shots is also possible. The interface of the search engine, depicted in Fig. 3, allows for easy query selection and swift visualization of results.

3 Experimental Setup

We performed our experiments as part of the interactive search task of the 2005 NIST TRECVID benchmark to demonstrate the significance of the proposed video search engine. The archive used is composed of 169 hours of US, Arabic, and Chinese broadcast news sources, recorded in MPEG-1 during November 2004. The test data contains approximately 85 hours. Together with the video archive came automatic speech recognition results and machine translations donated by a US government contractor. The Fraunhofer Institute [16] provided a camera shot segmentation. The camera shots serve as the unit for retrieval.

We detect in this data set automatically an unprecedented lexicon of 101 concepts using the semantic pathfinder. We select concepts by following a predefined concept ontology for multimedia [17] as leading example. Concepts in this ontology are chosen

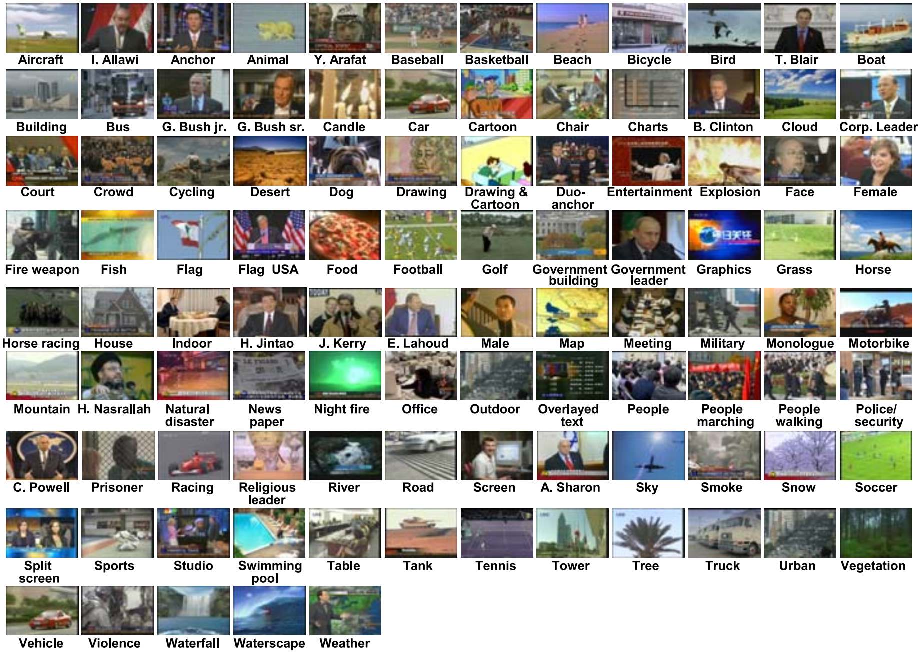

Fig. 4. Instances of the 101 concepts in the lexicon, as detected with the semantic pathfinder

based on presence in WordNet [18] and extensive analysis of video archive query logs. Where concepts should be related to program categories, setting, people, objects, activities, events, and graphics. Instantiations of the concepts in the lexicon are visualized in Fig. 4. The semantic pathfinder detects all 101 concepts with varying performance, see [8] for details.

The goal of the interactive search task, as defined by TRECVID, is to satisfy an information need. Given such a need, in the form of a search topic, a user is engaged in an interactive session with a video search engine. Based on the results obtained, a user rephrases queries; aiming at retrieval of more and more accurate results. To limit the amount of user interaction and to measure search system efficiency, all individual search topics are bounded by a 15-minute time limit. The interactive search task contains 24 search topics in total. They became known only few days before the deadline of submission. Hence, they were unknown at the time we developed our 101 semantic concept detectors. In line with the TRECVID submission procedure, a user was allowed to submit, for assessment by NIST, up to a maximum of 1,000 ranked results for the 24 search topics.

We use *average precision* to determine the retrieval accuracy on individual search topics, following the standard in TRECVID evaluations [12]. The average precision is a single-valued measure that is proportional to the area under a recall-precision curve. As an indicator for overall search system quality, TRECVID reports the mean average precision averaged over all search topics from one run by a single user.

4 Results

The complete numbered list of search topics is plotted in Fig. 5. Together with the topics, we plot the benchmark results for 49 users using 16 present-day interactive video search engines. We remark that most of them exploit only a limited lexicon of concepts, typically in the range of 0 to 40. The results give insight in the contribution of the proposed system for individual search topics. At the same time, it allows for comparison against the state-of-the-art in video retrieval.

The user of the proposed search engine scores excellent for most search topics, yielding a top 3 average precision for 17 out of 24 topics. Furthermore, our approach obtains the highest average precision for five search topics (Topics: 3, 8, 10, 13, 20). We explain the success of our search engine, in part, by the lexicon used. In our lexicon, there was an (accidental) overlap with the requested concepts for most search topics. Examples are *tennis*, *people marching*, and *road* (Topics: 8, 13, 20), where performance is very good. The search engine performed moderate for topics that require specific instances of a concept, e.g. maps with Bagdhad marked (Topic: 7). When search topics contain combinations of several concepts, e.g. meeting, table, people (Topic: 15), results are also not

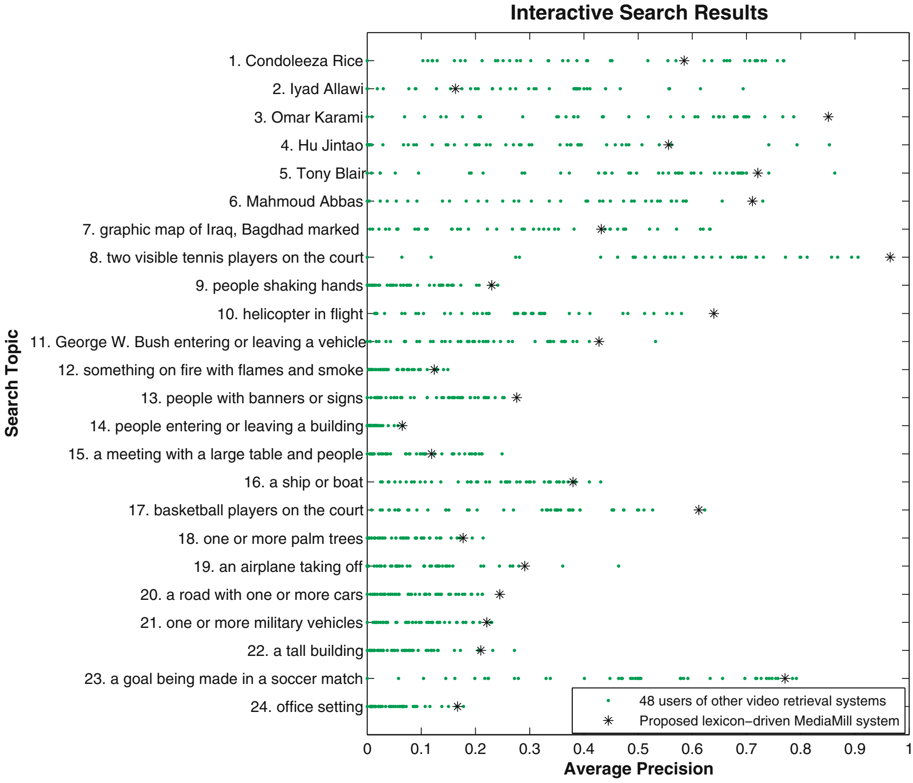

Fig. 5. Comparison of interactive search results for 24 topics performed by 49 users of 16 present-day video search engines

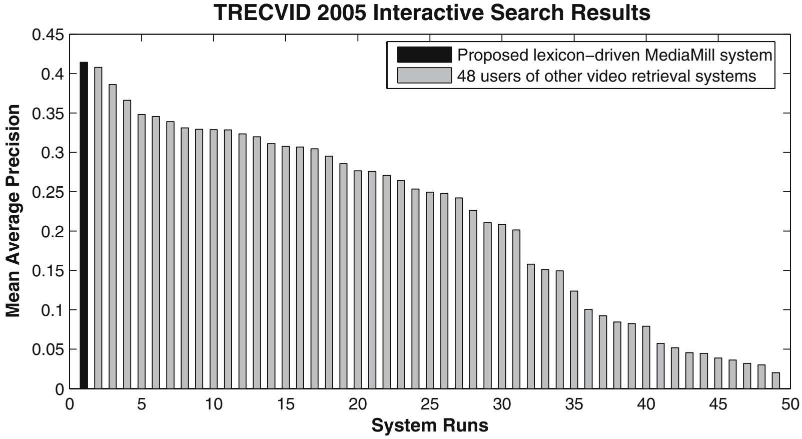

Fig. 6. Overview of all interactive search runs submitted to TRECVID 2005, ranked according to mean average precision

optimal. This indicates that much is to be expected from a more intelligent combination of query results. When a user finds an answer to a search topic in a repeating piece of footage, query-by-example is particularly useful. A typical search topic profiting from this observation it the one related to Omar Karami (Topic: 3), who is frequently interviewed in the same room. Query-by-keyword is especially useful for specific information needs, like person X related inquiries. It should be noted that although we have a large lexicon of concepts, performance of them is far from perfect, often resulting in noisy detection results. We therefore grant an important role to the interface of the video search engine. Because our user could quickly select relevant segments of interest, the search engine aided for search topics that could not be addressed with (robust) concepts from the lexicon.

To gain insight in the overall quality of our lexicon-driven approach to video retrieval, we compare the mean average precision results of using our search engine with 48 other users that participated in the interactive retrieval task of the 2005 TRECVID benchmark. We visualize the results for all submitted interactive search runs in Fig. 6. The results show that the proposed search engine obtains a mean average precision of 0.414, which is the highest overall score. The benchmark results demonstrate that lexicon-driven interactive retrieval yields state-of-the-art accuracy.

5 Conclusion

In this paper, we combine automatic learning of a large lexicon of semantic concepts with traditional video retrieval methods into a novel approach to narrow the semantic gap. The foundation of the proposed approach is formed by a learned lexicon of 101 semantic concepts. Based on this lexicon, query-by-concept offers users a semantic entrance to video repositories. In addition, users are provided with an entry in the form of textual query-by-keyword and visual query-by-example. Interaction with the various query interfaces is handled by an advanced display of results, which provides feedback in the form of a cross browser. The resulting *MediaMill* semantic video search engine limits the influence of the semantic gap.

Experiments with 24 search topics and 85 hours of international broadcast news video indicate that the lexicon of concepts aids substantially in interactive search performance. This is best demonstrated in a comparison among 49 users of 16 present-day retrieval systems, none of them using a lexicon of 101 concepts, within the interactive search task of the 2005 NIST TRECVID video retrieval benchmark. In this comparison, the user of the lexicon-driven search engine gained the highest overall score.

References

1. Flickner, M., et al.: Query by image and video content: The QBIC system. IEEE Computer **28**(9) (1995) 23–32
2. Chang, S.F., Chen, W., Men, H., Sundaram, H., Zhong, D.: A fully automated content-based video search engine supporting spatio-temporal queries. IEEE TCSVT **8**(5) (1998) 602–615
3. Rui, Y., Huang, T., Ortega, M., Mehrotra, S.: Relevance feedback: A power tool in interactive content-based image retrieval. IEEE TCSVT **8**(5) (1998) 644–655
4. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content based image retrieval at the end of the early years. IEEE TPAMI **22**(12) (2000) 1349–1380
5. Naphade, M., Huang, T.: A probabilistic framework for semantic video indexing, filtering, and retrieval. IEEE Trans. Multimedia **3**(1) (2001) 141–151
6. Amir, A., et al.: IBM research TRECVID-2003 video retrieval system. In: Proc. TRECVID Workshop, Gaithersburg, USA (2003)
7. Snoek, C., Worring, M., Geusebroek, J., Koelma, D., Seinstra, F., Smeulders, A.: The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. IEEE TPAMI (2006) in press.
8. Snoek, C., et al.: The MediaMill TRECVID 2005 semantic video search engine. In: Proc. TRECVID Workshop, Gaithersburg, USA (2005)
9. Rautiainen, M., Ojala, T., Seppänen, T.: Analysing the performance of visual, concept and text features in content-based video retrieval. In: ACM MIR, NY, USA (2004) 197–204
10. Christel, M., Huang, C., Moraveji, N., Papernick, N.: Exploiting multiple modalities for interactive video retrieval. In: IEEE ICASSP. Volume 3., Montreal, CA (2004) 1032–1035
11. Adcock, J., Cooper, M., Girgensohn, A., Wilcox, L.: Interactive video search using multilevel indexing. In: CIVR. Volume 3569 of LNCS., Springer-Verlag (2005) 205–214
12. Smeaton, A.: Large scale evaluations of multimedia information retrieval: The TRECVid experience. In: CIVR. Volume 3569 of LNCS., Springer-Verlag (2005) 19–27
13. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New York, USA (1983)
14. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. J. American Soc. Inform. Sci. **41**(6) (1990) 391–407
15. Lee, J.: Analysis of multiple evidence combination. In: ACM SIGIR. (1997) 267–276
16. Petersohn, C.: Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In: Proc. TRECVID Workshop, Gaithersburg, USA (2004)
17. Naphade, et al.: A light scale concept ontology for multimedia understanding for TRECVID 2005. Technical Report RC23612, IBM T.J. Watson Research Center (2005)
18. Fellbaum, C., ed.: WordNet: an electronic lexical database. The MIT Press, Cambridge, USA (1998)