A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval

Cees G. M. Snoek, *Member, IEEE*, Marcel Worring, *Member, IEEE*, Dennis C. Koelma, and Arnold W. M. Smeulders, *Member, IEEE*

***Abstract*—Effective video retrieval is the result of an interplay between interactive query selection, advanced visualization of results, and a goal-oriented human user. Traditional interactive video retrieval approaches emphasize paradigms, such as query-by-keyword and query-by-example, to aid the user in the search for relevant footage. However, recent results in automatic indexing indicate that query-by-concept is becoming a viable resource for interactive retrieval also. We propose in this paper a new video retrieval paradigm. The core of the paradigm is formed by first detecting a large lexicon of semantic concepts. From there, we combine query-by-concept, query-by-example, query-by-keyword, and user interaction into the *MediaMill* semantic video search engine. To measure the impact of increasing lexicon size on interactive video retrieval performance, we performed two experiments against the 2004 and 2005 NIST TRECVID benchmarks, using lexicons containing 32 and 101 concepts, respectively. The results suggest that from all factors that play a role in interactive retrieval, a large lexicon of semantic concepts matters most. Indeed, by exploiting large lexicons, many video search questions are solvable without using query-by-keyword and query-by-example. In addition, we show that the lexicon-driven search engine outperforms all state-of-the-art video retrieval systems in both TRECVID 2004 and 2005.**

***Index Terms*—Benchmarking, concept learning, content analysis and indexing, interactive systems, multimedia information systems, video retrieval.**

I. Introduction

THE technology for searching through text has evolved to a mature level of performance. Browsers and search engines have found in the Internet a medium to prosper, opening new ways to do business, science, and to be social. All of this was realized in just 15 years. That success has whet the appetite for retrieval of multimedia sources, specifically of the medium video. Present-day commercial video search engines [1], [2] often rely on just a filename and text metadata in the form of closed captions [1] or transcribed speech [2]. This results in a disappointing performance, as quite often the visual content is not mentioned, or properly reflected in the associated text. The text often covers the emotion of the video, but this is highly specific for context and wears quickly. In addition, when videos originate from non-English speaking countries, such as China or the Netherlands, querying the content becomes more difficult because automatic speech recognition is so much harder to achieve. At any rate, visual analysis up to the standards of text will deliver robustness to the multimedia search.

In contrast to text-based video retrieval, the content-based image retrieval research community has emphasized a visual-only approach. It has resulted in a wide variety of image and video search systems [3]–[13]. A common denominator in these prototypes is their dependence on low-level visual information such as color, texture, shape, and spatiotemporal features. Users query an archive containing visual feature values rather than the images. They do so by sketches, or by providing example images using a browser interface. Query-by-example can be fruitful when users search for the same object under slightly varying circumstances and when the target images are available indeed. If proper example images are unavailable, content-based image retrieval techniques are not effective at all. Moreover, users often do not understand similarity of low-level visual features. They expect semantic similarity. In other words, when searching for cars, an input image of a red car should also trigger the retrieval of yellow colored cars. The current generation of video search engines offers low-level abstractions of the data, where users seek high-level semantics. Thus, query-by-example retrieval techniques are not that effective in fulfilling the users' needs. The main problem for any video retrieval methodology aiming for access is the semantic gap between image data representation and their interpretation by humans [14]. Not surprisingly, the user experience with (visual only) video retrieval is one of frustration. Therefore, a new paradigm of semantics is required when aiming for access to video archives.

In a quest to narrow the semantic gap, recent research efforts have concentrated on automatic detection of semantic concepts in video. The feasibility of mapping low-level (visual) features to high-level concepts was proven by pioneering work, which distinguished between concepts such as *indoor* and *outdoor* [15], and *cityscape* and *landscape* [16]. The introduction of multimedia analysis, coupled with machine learning, has paved the way for generic indexing approaches [17]–[25]. Currently yielding concept lexicons bounded by 101 concepts [25], and expected to evolve into multimedia ontologies [26] containing as much as 1000 concepts soon [27]. The speed at which these lexicons grow offers great potential for future video retrieval systems. At present, the lexicons are not large enough, so they are no alternative yet for either the visual or textual retrieval paradigm. However, the availability of gradually increasing concept lexicons, raises the question: how to augment query-by-concept for effective interactive video retrieval?

Manuscript received April 2, 2006; revised June 28, 2006. This work was supported by the BSIK MultimediaN project. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Anna Hac.

The authors are with the Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, 1098 SJ Amsterdam, The Netherlands (e-mail: cgmsnoek@science.uva.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2006.886275

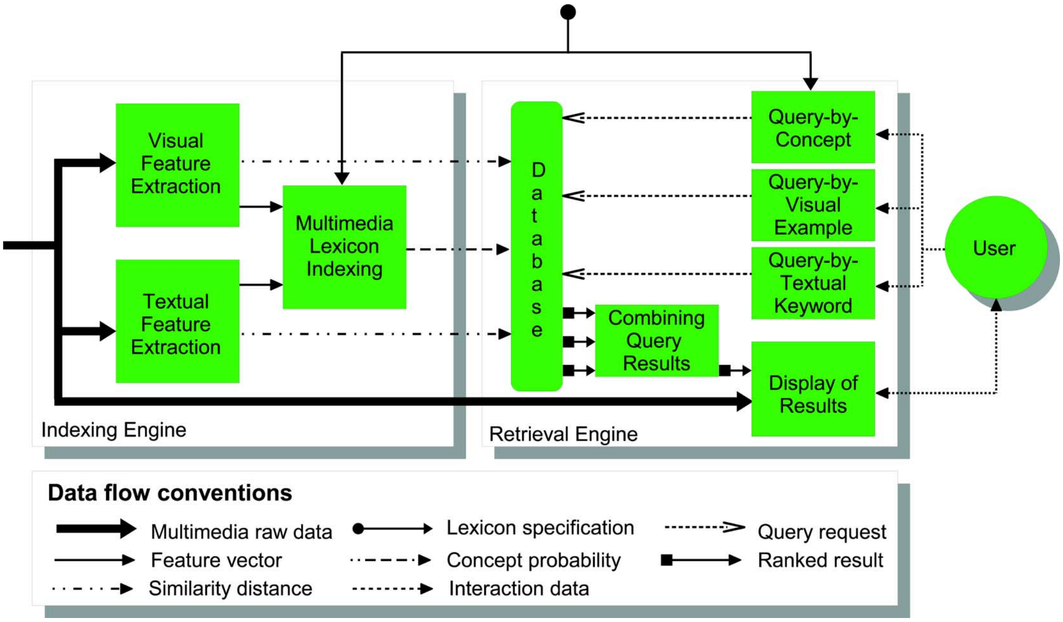


Fig. 1. General framework for an interactive video search engine. In the indexing engine, the system learns to detect a lexicon of semantic concepts. In addition, it computes similarity distances. A retrieval engine then allows for several query selection methods. The system combines requests and displays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

We start from the premise that a video search engine should begin with off-line learning of a large lexicon of multimedia concepts. In order to be effective in its use, a video search engine should employ query-by-example, query-by-keyword, and interaction with an advanced user interface to refine the search until satisfaction. We propose a *lexicon-driven paradigm* to video retrieval. The uniqueness of the proposed paradigm lies in its emphasis on automatic learning of a large lexicon of concepts. When the 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, which we call the *MediaMill* semantic video search engine. To demonstrate the effectiveness of our lexicon-driven retrieval paradigm, the interactive search experiments with the MediaMill system are evaluated within the 2004 and 2005 NIST TRECVID video retrieval benchmark [28], [29].

The organization of this paper is as follows. First, we formulate the problem in terms of related work in Section II. The blueprint of our lexicon-driven video retrieval paradigm is presented in Section III, where we describe the MediaMill system. We present the experimental setup in which we evaluated our paradigm in Section IV. We show the results of our experiments in Section V.

II. Problem Formulation and Related Work

We aim at providing users with semantic access to video archives. Specifically, we investigate whether this can be reached by machine learning. Then the question is how this large lexicon of learned concepts can be combined with query-by-keyword, query-by-example, and interactive manipulation to achieve effective video retrieval? In response to this question, we focus on methodologies that advocate the combination of lexicon learning, query-by-example, query-by-keyword, and interaction for semantic access [19], [20], [30], [31], [32]. We observe that these state-of-the-art video search systems 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. This similarity index facilitates retrieval in the form of query-by-example and query-by-keyword. Video search engines often employ a semantic indexing component to learn a lexicon of concepts 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. The system has an interface to compose queries, e.g., using query-by-keyword, query-by-example, and query-by-concept. The retrieval engine handles the query requests, combines the results, and displays them to an interacting user. A general framework for interactive video search engines is presented in Fig. 1. While proposed solutions for effective video search engines share similar components, they stress different elements in reaching their goal.

The interactive video retrieval system proposed by Adcock *et al.* [30] combines textual analysis with an advanced user interface. Their textual analysis automatically segments recognized speech transcripts on a topic-based story level. They argue that search results should be presented in these semantically meaningful story units. Therefore, they present query-by-keyword results as story key frame collages in the user interface. Their system does not support query-by-example and query-by-concept.

In contrast to [30], Taskiran *et al.* [31] stress visual analysis for retrieval, in particular similarity of low-level color features.

TABLE I
OVERVIEW OF STATE-OF-THE-ART VIDEO RETRIEVAL SYSTEMS, THEIR KEY-COMPONENTS, AND EVALUATION DETAILS, SORTED BY LEXICON SIZE. OUR CONTRIBUTION IS DENOTED IN BOLD

Reference	*Query-by-keyword*	*Query-by-example*	*Query-by-concept*	*Lexicon size*	*Display of results*	*Evaluation*
Adcock *et al.* [30]	✓			0	Story board	TRECVID 2004
Taskiran *et al.* [31]		✓	✓	1	Similarity pyramid	Specific
Fan *et al.* [20]		✓	✓	5	Hierarchical summarization	Specific
Christel *et al.* [32]	✓	✓	✓	10	Story board	TRECVID 2003
Amir *et al.* [19]	✓	✓	✓	17	Grid browser	TRECVID 2003
Snoek *et al.*	✓	✓	✓	**32**	**Grid browser**	**TRECVID 2004**
Snoek *et al.*	✓	✓	✓	**101**	**Cross browser**	**TRECVID 2005**

In addition, the authors provide users with a lexicon containing 1 concept, namely *face*. Obviously, a single concept can never address a wide variety of search topics. Thus, user interaction with the data is required. To that end, segmented shots are represented as an hierarchy of clustered frames. The authors combine this representation with query-by-example and query-by-concept by offering users query results in a so-called similarity pyramid. While users browse through the pyramid, they are offered a sense of the video archive at various levels of (visual) detail. Unfortunately its effectiveness remains unclear, as a verification on interactive retrieval experiments is missing.

In addition to visual analysis, Fan *et al.* [20] emphasize the utility of a lexicon, containing five concepts, for video retrieval. The authors exploit a hierarchical classifier to index the video on shot, scene, and cluster level, allowing for hierarchical browsing of video archives on concept-level and visual similarity. Unfortunately, similar to [31], the paper lacks an evaluation of the utility of the proposed framework for interactive video retrieval. A lexicon of five concepts aids for interactive video retrieval, but is still limited.

One of the first systems to combine query-by-keyword, query-by-example, query-by-concept, and advanced display of results is the Informedia system [32]–[34]. It is especially strong in interactive search scenarios. In [32], the authors explain the success in interactive retrieval as a consequence of using storyboards, i.e., a grid of key frame results that are related to a keyword-based query. As queries for semantic concepts are hard to tackle using the textual modality only, the interface also supports filtering based on semantic concepts. The filters are based on a lexicon of ten pre-indexed concepts with mixed performance [34]. Because the lexicon is limited in terms of the number of concepts, the filters are applied after a keyword-based search. 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. To limit the dependence on keywords, we emphasize query-by-concept in the interactive video retrieval process, where possible.

A system for generic semantic indexing is proposed by Naphade *et al.* in [17]–[19]. The system exploits consecutive aggregations on features, multiple modalities, and concepts. Finally, the system optimizes the result by rule-based post filtering. They report good benchmark results on a lexicon of 17 concepts. In spite of the use of this lexicon, interactive retrieval results with the web-driven *MARVEL* system [19] are not competitive with [30], [32]. This is surprising, given the robustness of the concept detectors. Hence, *MARVEL* has difficulty in properly leveraging the concept detection results for interactive retrieval. A drawback of the interactive system is the lack of speed of the web-based grid browser. Moreover, it has no video playback functionality. However, the largest problem is the complex query interface that offers too many possibilities to query on low-level (visual) features and prevents users from quick retrieval of video segments of interest. We adopt and extend their ideas related to semantic video indexing, but we take a different road for interactive retrieval.

From the need to quantify effective video retrieval, we note that it has always been a delicate issue. Video archives are fragmented and mostly inaccessible due to copyrights and the sheer volume of data involved. As a consequence, many researchers evaluate their video retrieval methodologies on specific data sets, e.g., [20], [31]. To make matters worse, as the evaluation requires substantial effort, they often evaluate submodules of the complete framework only. This is hampering progress because methodologies can not be valued on their relative merit with respect to interactive video retrieval performance. To tackle the evaluation problem, the American National Institute of Standards and Technology (NIST) started organizing the TRECVID video retrieval benchmark. The benchmark aims to promote progress in video retrieval via open, metrics-based evaluation [28], [29]. TRECVID provides video archives, a common shot segmentation, speech transcripts, and search topics that need to be solved by benchmark participants. Finally, they perform an independent examination of results using standard information retrieval evaluation measures. Because of its widespread acceptance in the field [28], [29], resulting in large participation of teams from academic labs, e.g., Carnegie Mellon University and Tsinghua University, and corporate research labs, e.g., IBM Research and FX Palo Alto Laboratory, the TRECVID benchmark can be regarded as the *de facto* standard to evaluate performance of video retrieval research.

To answer the questions related to combining video retrieval techniques and their joint evaluation in an interactive video retrieval setting, we first summarize our analysis of related work in Table I. It shows that interactive video retrieval methodologies stress different components indeed. We argue that a large lexicon of concepts matters most, i.e., query-by-concept should receive more emphasis in favor of traditional retrieval techniques. In this paper, we propose a lexicon-driven retrieval paradigm to equip users with semantic access to video archives (denoted in bold in Table I). The paradigm combines learning of a large lexicon—currently containing 32 concepts and 101 concepts, respectively—with query-by-keyword, query-by-example, and interaction using an advanced display of results. We introduce the MediaMill semantic video search engine, which exploits a grid browser and a cross browser for display of results, to demonstrate the effectiveness of the proposed paradigm. Since the search engine combines several techniques, we will not discuss in-depth technical details of individual components, nor will we evaluate them. In contrast, we focus on the performance of the combined approach to interactive video retrieval using accepted benchmarks. To that end, we evaluate our lexicon-driven retrieval paradigm within the 2004 and 2005 NIST TRECVID benchmark. Interactive retrieval using the proposed paradigm facilitates effective and efficient semantic access to video archives.

III. Lexicon-Driven Retrieval With the Mediamill Semantic Video Search Engine

With the MediaMill search engine we aim 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, where needed we provide pointers to published papers covering in-depth technical details.

A. Indexing Engine

1) Textual & Visual Feature Extraction: To arrive at a similarity distance for the textual modality, we first derive words from automatic speech recognition results, obtained with standard tools, e.g., [35]. We exploit standard machine translation tools [36] in case the videos originate from non-English speaking countries. This allows for a generic approach. We remove common stop words from the English text using the SMART's English stop list [37]. We then construct a high dimensional vector space based on all remaining transcribed words. We rely on latent semantic indexing [38] 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 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 [39] using 32 bins for each color channel. Users compare key frames with Euclidean histogram distance.

2) Multimedia Lexicon Indexing: Generic semantic video indexing is required to obtain a large concept lexicon. In literature, several approaches are proposed [17]–[25]. The utility of supervised learning in combination with multimedia content analysis has proven to be successful, with recent extensions to include video production style [22] and the insight that concepts often co-occur in context [17]–[19]. We combine these successful approaches into an integrated video indexing architecture.

The design principle of our architecture is derived from 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 semantic indexing must be to reverse this authoring process, for which we proposed the semantic pathfinder [24], [25]. 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. An in-depth discussion of the various techniques used is beyond the scope of this paper. We restrict ourselves here to a summary of the semantic pathfinder.

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. For visual feature extraction, we use Gaussian color measurements following [40] where it is argued that the opponent color system is the best orthogonal color representation in a three number system. Additional smoothing of these values with Gaussian spatial (derivative) filters suppresses acquisition and compression noise. We vary the size of the Gaussian filters to obtain a color-texture representation robust for variations in the target size. We apply normalization of each opponent color derivative by its intensity to suppress global intensity variations. Finally, we obtain rotationally invariant features by combining Gaussian derivative filter responses into two chromatic gradients—see [24] for details. Based on these color-texture measurements, the procedure segments a key frame in terms of regional visual concepts, like *concrete*, *sand*, *skin*, and so on. The percentage of pixels associated to each of the visual concepts is used as a visual content vector. In the textual modality, we learn the relation between uttered speech and concepts, where we obtain English-specific text from automatic speech transcription and machine translation. We connect words to shots and derive a lexicon of uttered words that co-occur with a concept. For each concept, we compare the text associated with each shot yielding a text vector containing a histogram of words. We fuse both the visual and textual vector using vector concatenation. In the learning phase, the semantic pathfinder applies a support vector machine [41] to learn concept probabilities. Here, we use the LIBSVM implementation [42] with radial basis function and Platt's conversion method [43] to achieve a probability rather than a margin. We obtain good parameter settings automatically, by using an iterative search on a large number of parameter combinations. We select the combination that yields the

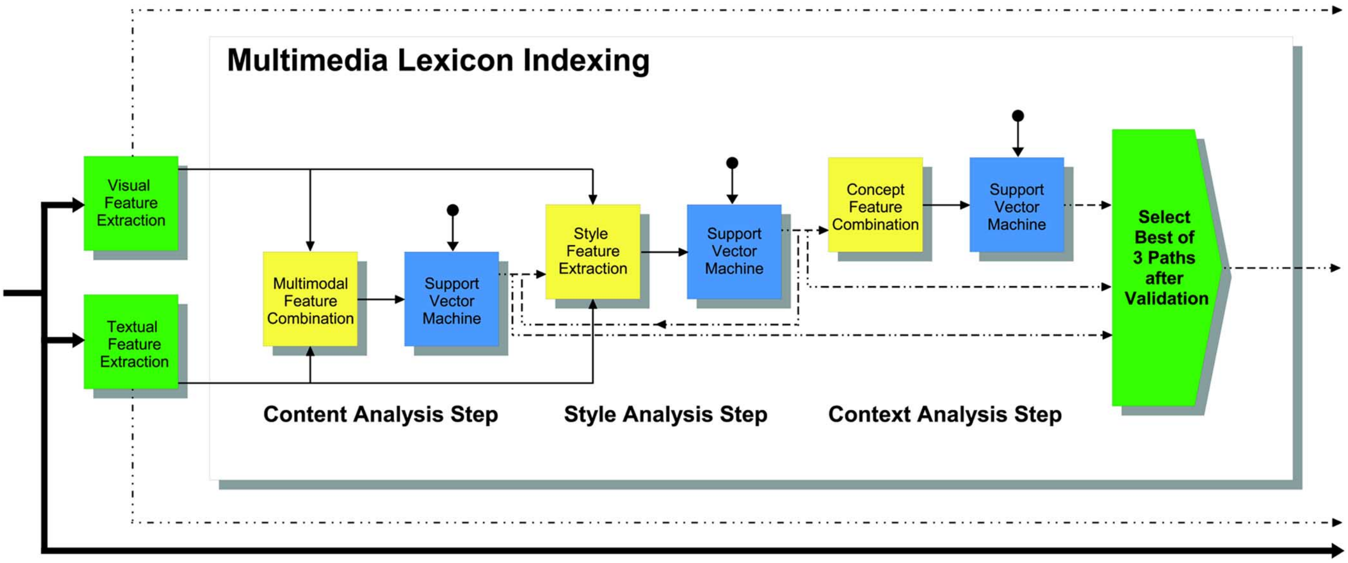

Multimedia Lexicon Indexing
Visual Feature Extraction
Textual Feature Extraction
Multimodal Feature Combination
Support Vector Machine
Style Feature Extraction
Support Vector Machine
Concept Feature Combination
Support Vector Machine
Select Best of 3 Paths after Validation
Content Analysis Step
Style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

Fig. 2. Multimedia lexicon indexing is based on the semantic pathfinder [24]. We highlight its successive analysis steps in the detail from Fig. 1. The semantic pathfinder selects for each concept a best analysis path after validation.

best retrieval performance after a threefold cross validation on training data.

In the *style analysis step*, we conceive of a video from the production perspective. Based on the four roles involved in the video production process [22], this step analyzes a video by four related style detectors. Layout detectors analyze the role of the editor, e.g., shot length, voice-over. Content detectors analyze the role of production design, e.g., face location, frequent speaker. Capture detectors analyze the role of the production recording unit, e.g., camera distance, camera motion. Finally, context detectors analyze the role of the preproduction team, e.g., studio, outdoor. We combine all style detector results using vector concatenation. Again, a support vector machine classifier is applied to learn optimized concept probabilities.

Finally, in the *context analysis step*, our aim is the reconstruction of the author's intent by considering detected concepts in context [17]. To that end, the probabilities obtained in the style analysis step are fused into a context vector. Then, a support vector machine classifier is again optimized to learn concepts.

The output of the context analysis step is also the output of the entire semantic pathfinder on video. On the way, we have included in the pathfinder, the results of the analysis on raw data, facts derived from production by the use of style features, and an intentional perspective of the author's objective by using concepts in context. For each concept we obtain a probability based on content, style, and context. The semantic pathfinder selects from the three possibilities the one that maximizes performance on validation data. It turns out that some concepts, like *vegetation*, have their emphasis on content thus style and context do not add much. An *interview* is a pure concept of style, where shots share many similarities in their production process while being very different content-wise. In contrast, more complex events, like *people walking*, profit from incremental adaptation of the analysis by using concepts like *athletic game* and *cityscape* 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.

B. Retrieval Engine

Video retrieval engines are often dictated by technical possibilities rather than actual user needs [44]. Frequently, this results in an overly complex system. 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 retrieval engine based on query selection methods. Each query method acts as a ranking operator on the video archive. After a user issues a query, it is processed and combined into a final result, which is presented to the user. The elements of our retrieval engine are now discussed in more detail.

1) Query Selection: The set of concepts in the lexicon forms the basis for interactive selection of query results. We identify three ways to exploit the lexicon for querying, i.e., *query-by-direct concept*, *query-by-subconcept*, and *query-by-super concept*. Users may rely on the query-by-direct concept for search topics related directly to concepts from the lexicon. In case the lexicon contains the concept *aircraft*, all information needs related to *aircrafts* benefit from query-by-direct concept. This is an enormous advantage for the precision of the search. Users can also make a first selection when a query includes a subconcept or a super-concept of a concept in the lexicon. For example, when searching for *sports* one can exploit query-by-subconcept using the available sport subconcepts *tennis*, *soccer*, *baseball*, and *golf* from the lexicon. In a similar fashion, users may exploit query-by-super concept using *animal* to retrieve footage related to *ice bear*. To aid the user in the selection of the query we make lexicon concepts available in the form of a subset of the WordNet [45] taxonomy. This helps the user to take well-established concept relations into account. The layout of the interface has the same order as WordNet for maximum comfort. In this

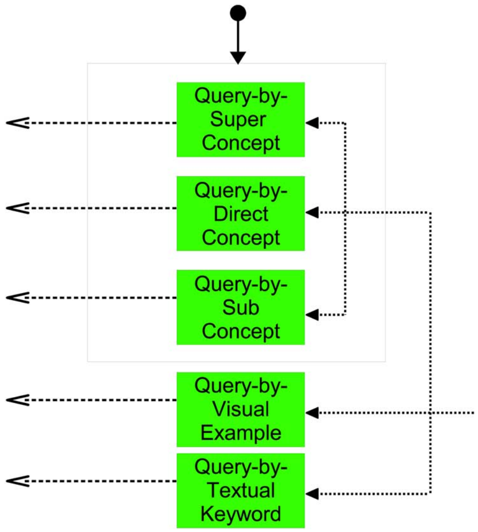


Fig. 3. MediaMill video search engine offers interacting users several methods for query selection. In the detail from Fig. 1, we highlight three query-by-concept methods, together with query-by-example, and query-by-keyword.

way, the lexicon of concepts aids users in various ways in specifying their queries.

For search topics not covered by the concepts in the lexicon, users have to rely on similarity distances in the form of query-by-keyword and query-by-example. Applying query-by-keyword in isolation allows users to find very specific topics only if they are mentioned in the transcription from automatic speech recognition. Based on query-by-example, on either provided or retrieved images, 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 provides users the possibility to combine query selection methods. 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 concept lexicon and the multimedia similarity distances. We summarize the methods for query selection in Fig. 3.

2) 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. As noted in Section II, one solution is to query the system in a sequential fashion. In such a scenario, a user may start with query-by-keyword, results obtained are subsequently filtered using query-by-concept. The disadvantage of this approach is the dependence on the accuracy of the initial query method. Therefore, we opt for a combination method that provides us the possibility to exploit query results in parallel. Rankings offer us a comparable output across various query results. Various ranking combination methods exist [46]. We employ a standard approach, using summation of linear rank normalizations [47], to combine query results.

3) Display of Results: The search engine supports two modes for displaying results. In the traditional *grid browser* a ranked list of key frame results is visualized as a lattice of thumbnails ordered left to right, top to bottom. However, ranking is a linear ordering. So, ideally it 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 is selected. This makes sense as frequently other items in the same broadcast are relevant to a query also [30], [32]. Therefore, we also employ a *cross browser*, which facilitates quick selection of relevant results. If requested, playback of specific shots is also possible. We rely on interaction by a user to select query methods and combine retrieval results. Technically, the interface of the search engine is implemented in *OpenGL* to allow for easy query selection and swift visualization of results. We depict the various aspects of the user interface of the MediaMill video search engine in Fig. 4.

IV. Experimental Setup

We investigate the impact of the proposed lexicon-driven paradigm for interactive video retrieval by performing two experiments with the MediaMill semantic video search engine.

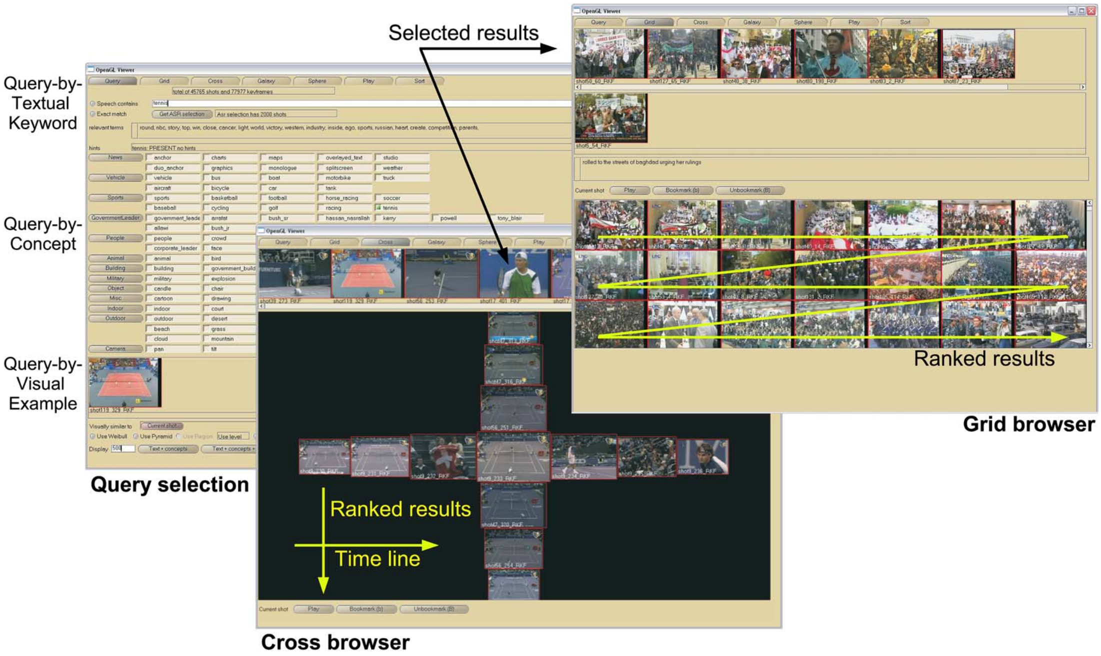
- **Experiment 1:** *Interactive video retrieval with a 32-concept lexicon*;
  In the first experiment, we evaluate video retrieval effectiveness using the MediaMill search engine in combination with a 32-concept lexicon and the grid browser.
- **Experiment 2:** *Interactive video retrieval with a 101-concept lexicon*;

In the second experiment, we evaluate video retrieval effectiveness using the MediaMill search engine in combination with a 101-concept lexicon and the cross browser. Finally, we compare interactive retrieval results obtained using the MediaMill search engine with a dozen other video retrieval systems. To allow for comparison, we perform all experiments as part of the interactive search tasks of the 2004 and 2005 NIST TRECVID benchmark.

A. Interactive Search

The goal of the interactive search task is to satisfy a number of video information needs. 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 min time limit. The 2004 interactive search task contains 23 search topics in total, the 2005 edition has 24. In line with the TRECVID submission procedure, a user was allowed to submit, for assessment by NIST, up to a maximum of 1000 ranked results for the various search topics.

The 2004 video archive includes 184 h of ABC World News Tonight and CNN Headline News. The training data contains approximately 120 h covering the period of January until June 1998. The test data holds the remaining 64 h, covering the period



total of 45765 shots and 77977 keyframes

○ Speech contains: tennis

○ Exact match

Asr selection has 2000 shots

relevant terms: round, nbc, story, top, win, close, cancer, light, world, victory, western, industry, inside, ago, sports, russian, heart, create, competition, parents,

hints: tennis: PRESENT no hints

☑ tennis

Visually similar to: Current shot

○ Use Weibull ○ Use Pyramid ○ Use Region Use level

Display: 500

Fig. 4. 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. For display of results, users may rely on a cross browser or a grid browser.

of October until December 1998. The 2005 archive contains 169 h with 287 episodes from 13 broadcast news shows from U.S., Arabic, and Chinese sources, recorded during November 2004. The test data contains approximately 85 h. Together with the video archives came automatic speech recognition results donated in 2004 by LIMSI [35] and in 2005 by a U.S. government contractor. CLIPS-IMAG [48] and the Fraunhofer Institute [49] provided a camera shot segmentation, in 2004 and 2005, respectively. The camera shots serve as the unit for retrieval.

B. Evaluation Criteria

To determine the retrieval accuracy on individual search topics TRECVID uses *average precision*. The average precision is a single-valued measure that is proportional to the area under a recall-precision curve. This value is the average of the precision over all relevant judged shots. Hence, it combines precision and recall into one performance value. To be precise, let $L^k = \{l_1, l_2, \dots, l_k\}$ be a ranked version of the answer set A. At any given rank k let $R \cap L^k$ be the number of relevant shots in the top k of L, where R is the total number of relevant shots. Average precision is defined by

$$\text{average precision} = \frac{1}{R} \sum_{k=1}^{A} \frac{R \cap L^k}{k} \lambda(l_k) \tag{1}$$

where indicator function $\lambda(l_k) = 1$ if $l_k \in R$ and 0 otherwise. As the denominator k and the value of $\lambda(l_k)$ are dominant in determining average precision, it can be understood that this metric favours highly ranked relevant shots.

TRECVID uses a pooled ground truth P, to reduce labor-intensive manual judgments of all submitted runs. They take from each submitted run a fixed number of ranked shots, which is combined into a list of unique shots. Every submission is then evaluated based on the results of assessing this merged subset, i.e., instead of using R in (1), P is used, where $P \subset R$. This yields an incomplete ground truth, but a fair comparison of submissions.

As an indicator for overall search system quality TRECVID computes the mean average precision over all search topics from one run by a single user.

C. Lexicon Specification

We automatically detect a lexicon of semantic concepts in both the TRECVID 2004 and 2005 data using the semantic pathfinder, as discussed in Section III-A2. In the 2004 data we detect a lexicon of 32 concepts, in the 2005 data lexicon of 101 concepts. We select concepts by following a predefined concept ontology for multimedia [50] as a leading example. Concepts in this ontology are chosen based on presence in WordNet [45] and extensive analysis of video archive query logs. Where concepts should be related to program categories, setting, people, objects, activities, events, and graphics. In addition, a primary design choice was that concepts need to be clear by looking at a static key frame only. We visualize instantiations of the detected concepts in both lexicons in Fig. 5, additional details for 2004 data are in [24] and for 2005 data are in [25].

It should be noted that although we have a large lexicon of concepts, with state-of-the-art results for generic indexing [24], performance of them is far from perfect. This often results in noisy detection outcomes. To give an indication of performance, we highlight our official TRECVID concept detection results on test data in Table II. The TRECVID procedure prescribes that ten pre-defined concepts are evaluated. Hence, for each year, we can report the official benchmark results for ten concepts in our lexicon only. The benchmark concepts are, however, representative for the entire lexicons.

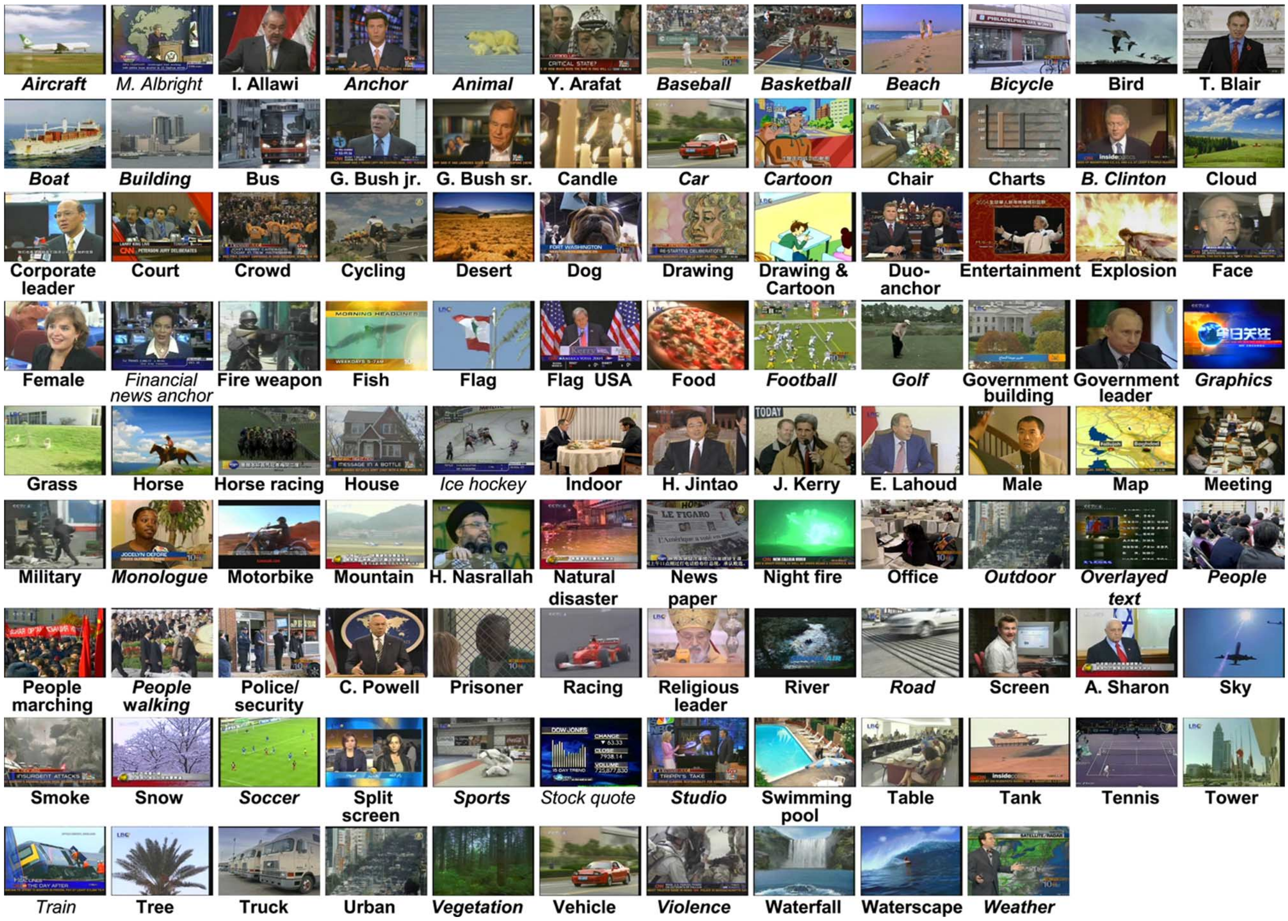


Fig. 5. Instances of the concepts from the lexicons used. The lexicon of 32 concepts (TRECVID 2004) is given in italics, the 101-concept lexicon (TRECVID 2005) is denoted in bold. Concepts which appear in both lexicons follow two conventions.

TABLE II
MEDIAMILL AVERAGE PRECISION RESULTS FOR THE TRECVID 2004 [24] AND 2005 [25] CONCEPT DETECTION TASK

TRECVID 2004		**TRECVID 2005**	
Concept	*Average Precision*	*Concept*	*Average Precision*
Aircraft	0.065	Building	0.235
M. Albright	0.136	Car	0.213
Basketball	0.209	Explosion	0.041
Beach	0.020	Flag USA	0.100
Boat	0.117	Map	0.142
B. Clinton	0.150	Mountain	0.220
People walking	0.170	People walking	0.199
Road	0.138	Prisoner	0.005
Train	0.062	Sports	0.342
Violence	0.086	Waterscape	0.201

We stress that the various topics became known only a few days before the deadline of submission. Hence, they were unknown at the time we developed our semantic concept detectors. Moreover, the test set performance of the concepts was unknown at the time we performed our interactive search experiments. To show the potential of our lexicon-driven paradigm we performed an experiment with a single expert user, which is common procedure in TRECVID, e.g., [30], [32] [19]. Our expert user had no experience with the topics nor with the test data. The user did have experience with the MediaMill system and its concept lexicons, but only on training data, which is conform TRECVID guidelines.

V. RESULTS

A. *Interactive Video Retrieval With a 32 Concept Lexicon*

We plot the complete numbered list of search topics used in our first experiment in Fig. 6. In addition, we plot the benchmark results for 61 users with 14 present-day interactive multimedia retrieval systems. The results give us insight in the contribution of the proposed paradigm for individual search topics when using a lexicon of 32 concepts.

For most search topics, the user of the proposed paradigm for interactive multimedia retrieval scores above average. Furthermore, the user of our approach obtains the highest average precision for seven search topics (Topics: 3, 14, 15, 16, 18, 20, 21). We explain the success of our interactive retrieval paradigm in this experiment in part by the lexicon used. In our lexicon, there was an (accidental) overlap with the requested concepts from some search topics; for example, *ice hockey*, *bicycle*, and *Bill Clinton* (Topics: 6, 16, 20), where performance is very good. Implying that there is much to be expected from a larger set of concepts in the lexicon. For other topics, the user could use query-by-super concept for filtering, e.g., *sporting event* for

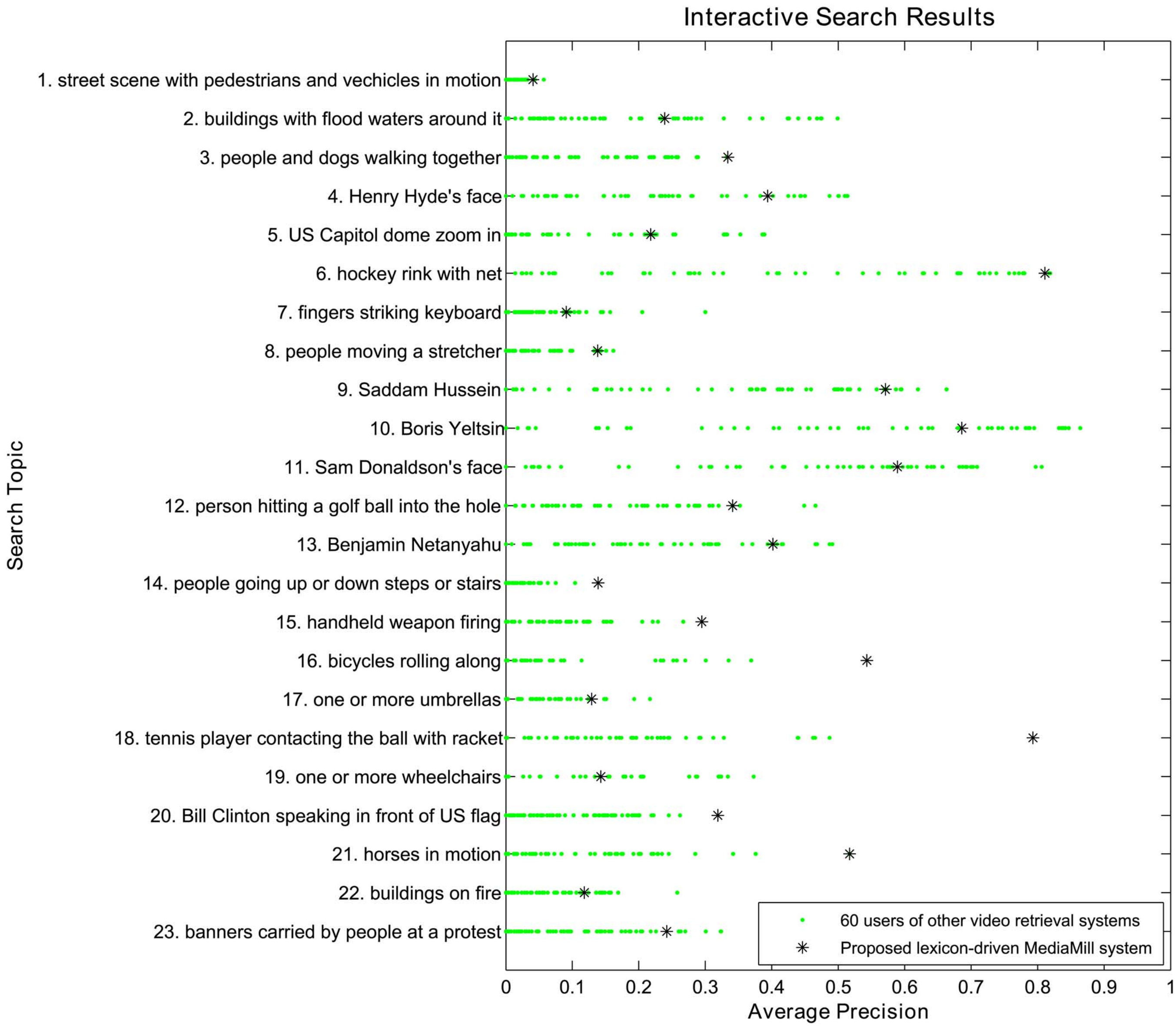


Fig. 6. Comparison of interactive search results for 23 topics performed by 61 users of 14 present-day video retrieval systems. Results for the user of the proposed paradigm, with a 32-concept lexicon, are indicated with special markers.

tennis player (Topic: 18) and *animal* for horses (Topic: 21). So in our method, abstract concepts make sense even when they are referred to indirectly. As an exception, for search topics related to the concept *building* (Topics: 2, 22), our retrieval method performed badly compared to the best results. We explain this behavior by the fact that building was not the distinguishing concept in these topics, but rather concepts like *flood* and *fire*, implying that some concepts are more important than others.

The user of the paradigm performed moderate for search topics that did not have a clear overlap with the concepts in the lexicon. For topics related to wheelchairs (Topic: 19), umbrellas (Topic: 17), and person X who were not in the lexicon (Topics: 4, 9, 10, 11, 13), query-by-keyword is the only viable alternative.

When a user recognizes an answer to a search topic as a commercial or signature tune, query-by-example is particularly useful. Search topics profiting from this observation are those related to bicycle and tennis player (Topics: 16, 18). Since these fragments contain similar visual content throughout the archive, they are easily retrievable with query-by-example.

After this first experiment, we conclude that for search topics related to concepts in the lexicon, query-by-concept is a good starting point. Query-by-keyword is effective when the (visual) content is described in the speech signal. If a user is interested in footage that is repeated throughout the archive, query-by-example is the way to go. With a lexicon containing only 32 concepts, we already diminish the influence of traditional video retrieval techniques in favor of query-by-concept.

B. *Interactive Video Retrieval With a 101 Concept Lexicon*

We again plot the complete numbered list of search topics in Fig. 7 for our second experiment, where we use a lexicon of 101 concepts. Together with the topics, we plot the benchmark

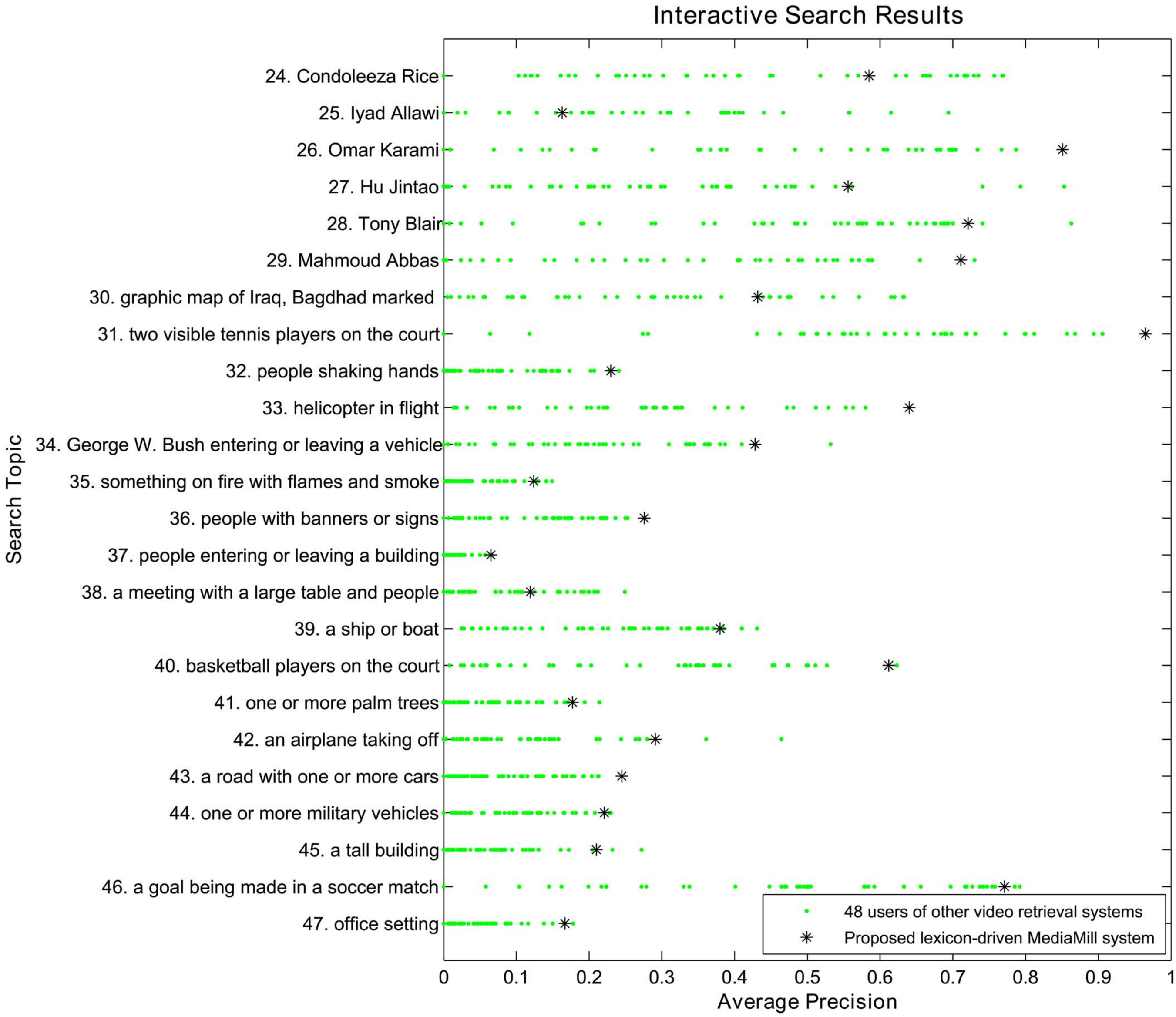


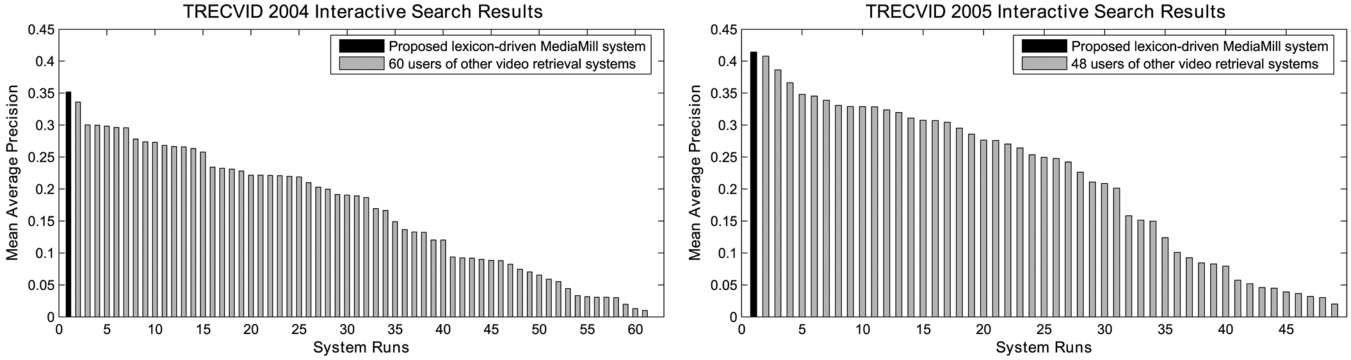
Fig. 7. Comparison of interactive search results for 24 topics performed by 49 users of 16 present-day video retrieval systems. Results for the user of the proposed paradigm, with a 101-concept lexicon, are indicated with special markers.

results for 49 users using 16 present-day interactive video search engines.

The results confirm the value of a large lexicon for interactive video retrieval. For most search topics the user of the proposed paradigm scores excellent, 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: 26, 31, 33, 36, 43). In our lexicon, there was an (accidental) overlap with the requested concepts from almost all search topics; for example *tennis*, *people marching*, and *road* (Topics: 31, 36, 43), where performance is very good. These results demonstrate that many video search questions are solvable without using query-by-keyword and query-by-example.

The search engine performed moderate for topics that were not in the lexicon (Topic: 24), or which yielded very poor concept detection (Topic: 25). For these topics, our user had to rely on query-by-keyword. In addition, we also performed less than expected for topics that require specific instances of a concept, e.g., maps with Bagdhad marked (Topic: 30). Although the concept *map* was part of our lexicon, our user was unable to exert this advantage. When search topics contain combinations of several reliable concepts, e.g., meeting, table, people (Topic: 38), results are also not optimal. This indicates that much is to be expected from a more intelligent combination of query results.

For some topics, the MediaMill search engine may be exploited in an unexpected way. By the use of common sense, the lexicon is also useful for topics that do not have a clear one-to-one relation with a concept. One of the search topics profiting from this observation is people shaking hands (Topic: 32). For this topic, the concept *government leader* is helpful. Indeed, government leaders shake hands quite often when visiting or welcoming fellow foreign leaders, which is often broadcasted in news items. For the topic on finding one or more palm trees (Topic: 41), query-by-direct concept on *tree* was not spe-



(a) (b)

Fig. 8. Overview of all interactive search runs submitted to (a) TRECVID 2004 and (b) TRECVID 2005, ranked according to mean average precision.

cific enough. Here our user exploited common sense by using the fact that by searching on *military* the system returns a lot of shots from the war in Iraq. Indeed often containing palm trees. Lebanese former prime minister Omar Karami (Topic: 26) was not included in our lexicon. For this topic we combine common sense with the cross browser. Omar Karami appears often in long interviews. Thus, when a single shot from such an interview is localized, the cross browser offers an opportunity to select a large amount of relevant shots easily. When users employ common sense, the lexicon-driven paradigm becomes even more powerful.

Our second experiment shows that a large lexicon is the most valuable resource for interactive video retrieval. With a lexicon of 101 concepts, almost all search topics are solvable directly, or indirectly, with good performance. Hence, the value of the lexicon-driven paradigm is evident. In fact, it diminishes the value of traditional techniques such as query-by-keyword and query-by-example to purely supportive tools for topics that can not be addressed by concepts from the lexicon. Using a large lexicon implies a paradigm shift for interactive video retrieval.

C. Benchmark Comparisons

To gain insight in the overall quality of our lexicon-driven interactive retrieval paradigm, we compare the mean average precision results of our lexicon-driven MediaMill video search engine with other state-of-the-art systems. For TRECVID 2004 we compare against 13 other retrieval systems. For TRECVID 2005 we compare against 15 present-day video search engines. Our approach is unique with respect to lexicon size, most others emphasize traditional retrieval paradigms. We visualize the results for all submitted interactive search runs of TRECVID 2004 in Fig. 8(a), and TRECVID 2005 in Fig. 8(b).

The results show that the proposed search engine obtains a mean average precision of 0.352 in TRECVID 2004, and 0.414 in TRECVID 2005. In both cases, the highest overall score. In [51] the authors showed that the top two TRECVID 2004 systems significantly outperform all other submissions. What is striking about these results, is that we obtain them by using a lexicon of only 32 concepts. When we increase the concept lexicon to 101 concepts in TRECVID 2005, only three users stay within a difference of 0.05 mean average precision. These users employed a video retrieval system based on rapid serial visual presentation of search results [52]. In such a scenario, a user is bombarded with as much key frames as possible. While effective in terms of TRECVID performance, this demanding approach is suited for limited domains only. The benchmark results demonstrate that lexicon-driven interactive retrieval yields superior performance relative to state-of-the-art video search engines.

VI. Conclusion

In this paper, we combine automatic learning of a large lexicon of semantic concepts with video retrieval methods into an effective video search system. The aim of the combined system is to narrow the semantic gap for the user. The foundation of the proposed approach is to learn a lexicon of semantic concepts. Where it should be noted that we have used a generic machine learning system and no per-concept optimizations. Based on this learned 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 grid browser or a cross browser. The resulting *MediaMill* semantic video search engine limits the influence of the semantic gap.

We investigate the impact of the proposed lexicon-driven paradigm for interactive video retrieval by performing two experiments with the MediaMill semantic video search engine. In our first experiment, with a lexicon of only 32 concepts, we already outperform state-of-the-art systems in seven out of 23 random queries on 64 h of U.S. broadcast news. When we increase the lexicon to 101 concepts, in our second experiment, we obtain a top three average precision for 17 out of 24 topics and top performance for five topics on an 85 h international news video archive. The key insight resulting from these experiments is that from all factors that play a role in interactive retrieval, a large lexicon of semantic concepts matters most. This is best demonstrated when we compare our lexicon-driven approach against the 2004 and 2005 NIST TRECVID benchmark. In both cases, our MediaMill system obtains superior performance relative to a dozen other state-of-the-art video search engines, which still adhere to traditional video retrieval paradigms.

Retrieval results with the proposed paradigm range from "poor" for topics like "find street scenes with pedestrians and vehicles in motion" to excellent for non-trivial topics like "find two visible tennis players on the court". However, under all topics, the performance is good relative to other systems and best overall. Fluctuating performance of multimedia retrieval technology is unacceptable for highly demanding applications, such as military intelligence. However, when used in a less demanding commercial search scenario, the proposed paradigm provides already valuable semantic information.

Acknowledgment

The authors are grateful to NIST and the TRECVID coordinators for the benchmark organization and evaluation effort.

References

[1] Google Video Search 2006 [Online]. Available: http://video.google.com/

[2] Blinkx Video Search 2006 [Online]. Available: http://www.blinkx.tv/

[3] T. Kato, T. Kurita, N. Otsu, and K. Hirata, "A sketch retrieval method for full color image database—query by visual example," in *Proc. Int. Conf. Pattern Recognition*, The Hague, The Netherlands, 1992, vol. 1, pp. 530–533.

[4] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *IEEE Comput.*, vol. 28, no. 9, pp. 23–32, Sep. 1995.

[5] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *Int. J. Comput. Vis.*, vol. 18, no. 3, pp. 233–254, 1996.

[6] A. Del Bimbo and P. Pala, "Visual image retrieval by elastic matching of user sketches," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 121–132, Feb. 1997.

[7] A. Gupta and R. Jain, "Visual information retrieval," *Commun. ACM*, vol. 40, no. 5, pp. 70–79, 1997.

[8] J. Smith and S.-F. Chang, "Visually searching the web for content," *IEEE Multimedia*, vol. 4, no. 3, pp. 12–20, Jul.–Sep. 1997.

[9] S.-F. Chang, W. Chen, H. Men, H. Sundaram, and D. Zhong, "A fully automated content-based video search engine supporting spatio-temporal queries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 602–615, Sep. 1998.

[10] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool in interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, Sep. 1998.

[11] W. Ma and B. Manjunath, "Netra: a toolbox for navigating large image databases," *Multimedia Syst.*, vol. 7, no. 3, pp. 184–198, 1999.

[12] T. Gevers and A. W. M. Smeulders, "Pictoseek: Combining color and shape invariant features for image retrieval," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 102–119, Jan. 2000.

[13] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1026–1038, Aug. 2002.

[14] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[15] M. Szummer and R. Picard, "Indoor-outdoor image classification," in *IEEE Int. Workshop Content-based Access Image Video Databases*, Bombay, India, 1998.

[16] A. Vailaya, M. Figueiredo, A. Jain, and H.-J. Zhang, "Image classification for content-based indexing," *IEEE Trans. Image Process.*, vol. 10, pp. 117–130, Jan. 2001.

[17] M. Naphade and T. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 141–151, Mar. 2001.

[18] W. H. Adams, G. Iyengar, C.-Y. Lin, M. Naphade, C. Neti, H. Nock, and J. Smith, "Semantic indexing of multimedia content using visual, audio, and text cues," *EURASIP J. Appl. Signal Process.*, no. 2, pp. 170–185, 2003.

[19] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M. Naphade, A. Natsev, C. Neti, H. Nock, J. Smith, B. Tseng, Y. Wu, and D. Zhang, "IBM research TRECVID-2003 video retrieval system," in *Proc. TRECVID Workshop*, Gaithersburg, MD, 2003.

[20] J. Fan, A. Elmagarmid, X. Zhu, W. Aref, and L. Wu, "ClassView: Hierarchical video shot classification, indexing, and accessing," *IEEE Trans. Multimedia*, vol. 6, no. 1, pp. 70–86, Feb. 2004.

[21] J. Fan, H. Luo, and A. Elmagarmid, "Concept-oriented indexing of video databases: Toward semantic sensitive retrieval and browsing," *IEEE Trans. Image Process.*, vol. 13, no. 7, pp. 974–992, Jul. 2004.

[22] C. Snoek, M. Worring, and A. Hauptmann, "Learning rich semantics from news video archives by style analysis," *ACM Trans. Multimedia Comput., Commun. Applic.*, vol. 2, no. 2, pp. 91–108, 2006.

[23] C. Snoek, M. Worring, J. Geusebroek, D. Koelma, and F. Seinstra, "The MediaMill TRECVID 2004 semantic video search engine," in *Proc. TRECVID Workshop*, Gaithersburg, MD, 2004.

[24] C. Snoek, M. Worring, J. Geusebroek, D. Koelma, F. Seinstra, and A. Smeulders, "The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1678–1689, Oct. 2006.

[25] C. Snoek, J. v. Gemert, J. Geusebroek, B. Huurnink, D. Koelma, G. Nguyen, O. d. Rooij, F. Seinstra, A. Smeulders, C. Veenman, and M. Worring, "The MediaMill TRECVID 2005 semantic video search engine," in *Proc. TRECVID Workshop*, Gaithersburg, MD, 2005.

[26] A. Benitez, J. Smith, and S.-F. Chang, "MediaNet: A multimedia information network for knowledge representation," in *Proc. SPIE Conf. Internet Multimedia Management Syst.*, Boston, MA, 2000, vol. 4210.

[27] A. Hauptmann, "Towards a large scale concept ontology for broadcast video," in *Conference on Image and Video Retrieval (CIVR)*, ser. LNCS. New York: Springer-Verlag, 2004, vol. 3115, pp. 674–675.

[28] A. Smeaton, P. Over, and W. Kraaij, "TRECVID: Evaluating the effectiveness of information retrieval tasks on digital video," in *ACM Multimedia*. New York: , 2004.

[29] A. Smeaton, "Large scale evaluations of multimedia information retrieval: The TRECVID experience," in *Conference on Image and Video Retrieval (CIVR)*, ser. LNCS. New York: Springer-Verlag, 2005, vol. 3568, pp. 19–27.

[30] J. Adcock, M. Cooper, A. Girgensohn, and L. Wilcox, "Interactive video search using multilevel indexing," in *Conference on Image and Video Retrieval (CIVR)*, ser. LNCS. New York: Springer-Verlag, 2005, vol. 3568, pp. 205–214.

[31] C. Taskiran, J.-Y. Chen, A. Albiol, L. Torres, C. Bouman, and E. Delp, "ViBE: a compressed video database structured for active browsing and search," *IEEE Trans. Multimedia*, vol. 6, no. 1, pp. 103–118, Feb. 2004.

[32] M. Christel, C. Huang, N. Moraveji, and N. Papernick, "Exploiting multiple modalities for interactive video retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Montreal, QC, Canada, 2004, vol. 3, pp. 1032–1035.

[33] H. Wactlar, M. Christel, Y. Gong, and A. Hauptmann, "Lessons learned from building a terabyte digital video library," *IEEE Comput.*, vol. 32, no. 2, pp. 66–73, 1999.

[34] A. Hauptmann, R. Baron, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, N. Moraveji, N. Papernick, C. Snoek, G. Tzanetakis, J. Yang, R. Yang, and H. Wactlar, "Informedia at TRECVID 2003: analyzing and searching broadcast news video," in *Proc. TRECVID Workshop*, Gaithersburg, MD, 2003.

[35] J. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Commun.*, vol. 37, no. 1–2, pp. 89–108, 2002.

[36] K. Knight and D. Marcu, "Machine translation in the year 2004," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Philadelphia, PA, 2005, vol. 5, pp. 965–968.

[37] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[38] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inform. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[39] T. Gevers and A. Smeulders, "Content-based image retrieval: An overview," in *Emerging Topics in Computer Vision*, G. Medioni and S. Kang, Eds. Upper Saddle River, NJ: Prentice-Hall, 2004.

[40] J. Geusebroek, R. v. d. Boomgaard, A. Smeulders, and H. Geerts, "Color invariance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1338–1350, Dec. 2001.

[41] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed ed. New York: Springer-Verlag, 2000.

[42] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines 2001 [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm/

[43] J. Platt, "Probabilities for SV machines," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 2000, pp. 61–74.

[44] H. Lee and A. Smeaton, "Designing the user-interface for the Físchlár digital video library," *J. Digital Inform.*, vol. 2, no. 4, pp. 251–262, 2002.

[45] C. Fellbaum, Ed., *WordNet: an electronic lexical database*. Cambridge, MA: The MIT Press, 1998.

[46] T. Ho, J. Hull, and S. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 1, pp. 66–75, Jan. 1994.

[47] J. Lee, "Analysis of multiple evidence combination," in *Proc. ACM SIGIR*, 1997, pp. 267–276.

[48] G. Quénot, D. Moraru, L. Besacier, and P. Mulhem, E. Voorhees and L. Buckl, Eds., "CLIPS at TREC-11: experiments in video retrieval," in *Proc. 11th Text REtrieval Conf.*, Gaithersburg, MD, 2002, vol. 500–251.

[49] C. Petersohn, "Fraunhofer HHI at TRECVID 2004: shot boundary detection system," in *Proc. TRECVID Workshop*, Gaithersburg, MD, 2004.

[50] M. Naphade, L. Kennedy, J. Kender, S.-F. Chang, J. Smith, P. Over, and A. Hauptmann, A light scale concept ontology for multimedia understanding for TRECVID 2005 IBM T. J. Watson Research Center, Yorktown Heights, NY, 2005, Tech. Rep. RC23612.

[51] A. Hauptmann and W.-H. Lin, "Assessing effectiveness in video retrieval," in *Conference on Image and Video Retrieval (CIVR)*, ser. LNCS. New York: Springer-Verlag, 2005, vol. 3568, pp. 215–225.

[52] A. Hauptmann, M. Christel, R. Concescu, J. Gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, and Y. Zhang, "CMU informedia's TRECVID 2005 skirmishes," in *Proc. TRECVID Workshop*, Gaithersburg, MD, 2005.

Cees G. M. Snoek (S'01–M'06) received the M.Sc. degree in business information systems in 2000 and the Ph.D. degree in computer science in 2005, both from the University of Amsterdam, Amsterdam, The Netherlands.

He is currently a Senior Researcher at the Intelligent Systems Lab, University of Amsterdam. He was a Visiting Scientist at Informedia, Carnegie Mellon University, Pittsburgh, PA, in 2003. His research interests focus on multimedia signal processing, statistical pattern recognition, content-based information retrieval, and large-scale benchmark evaluations, especially when applied in combination for multimedia understanding.

Dr. Snoek is the Local Chair of the 2007 ACM International Conference on Image and Video Retrieval in Amsterdam. He is a Lead Architect of the MediaMill video search engine, which obtained state-of-the-art performance in recent NIST TRECVID evaluations and was awarded as best technical demonstration at ACM Multimedia 2005.

Marcel Worring (M'03) received the M.Sc. (Hons.) degree from the Vrije Universiteit, Amsterdam, The Netherlands, in 1983 and the Ph.D. degree from the University of Amsterdam in 1998, both in computer science.

He is currently an Associate Professor at the University of Amsterdam. His interests are in multimedia searches and systems. He leads several multidisciplinary projects covering knowledge engineering, pattern recognition, image and video analysis, and information space interaction, conducted in close cooperation with industry. In 1998, he was a Visiting Research Fellow at the University of California, San Diego. He has published over 100 scientific papers and serves on the program committee of several international conferences.

Dr. Worring is the Chair of the IAPR TC12 on Multimedia and Visual Information Systems and the General Chair of the 2007 ACM International Conference on Image and Video Retrieval.

Dennis C. Koelma received the M.Sc. and Ph.D. degrees in computer science from the University of Amsterdam, Amsterdam, The Netherlands, in 1989 and 1996, respectively. His thesis was on the topic of a software environment for image interpretation.

Currently, he is working on Horus, a software architecture for doing research in accessing the content of digital images and video. His research interests include image and video processing, software architectures, parallel programming, databases, graphical user interfaces, and image information systems.

Arnold W. M. Smeulders (M'79) received the M.Sc. degree in physics from the Technical University of Delft, Delft, The Netherlands, in 1977 and the Ph.D. degree in medicine from Leiden University, Leiden, The Netherlands, in 1982, on the topic of visual pattern analysis.

He is the Scientific Director of the Intelligent Systems Lab, Amsterdam, The Netherlands, of the MultimediaN Dutch public-private partnership, and of the ASCI National Research School. His work in the Intelligent Sensory Information Systems (ISIS) group, as part of the Intelligent Systems Lab, concentrates on theory, practice, and implementation of multimedia information analysis, including image databases and computer vision; the group has an extensive record in co-operations with Dutch institutions and industry in the area of multimedia and video analysis. He participates in the EU-Vision, DELOS, and MUSCLE networks of excellence. His research interests include cognitive vision, content-based image retrieval, learning and tracking, and the picture-language question. He has written 300 papers in refereed journals and conferences and graduated 28 Ph.D. students.

Dr. Smeulders is an Associate Editor of the *International Journal of Computer Vision* and the IEEE TRANSACTIONS ON MULTIMEDIA. He is Fellow of the International Association of Pattern Recognition.