Adding Semantics to Detectors for Video Retrieval

Cees G. M. Snoek, *Member, IEEE*, Bouke Huurnink, Laura Hollink, Maarten de Rijke, Guus Schreiber, and Marcel Worring, *Member, IEEE*

***Abstract*—In this paper, we propose an automatic video retrieval method based on high-level concept detectors. Research in video analysis has reached the point where over 100 concept detectors can be learned in a generic fashion, albeit with mixed performance. Such a set of detectors is very small still compared to ontologies aiming to capture the full vocabulary a user has. We aim to throw a bridge between the two fields by building a multimedia thesaurus, i.e., a set of machine learned concept detectors that is enriched with semantic descriptions and semantic structure obtained from WordNet. Given a multimodal user query, we identify three strategies to select a relevant detector from this thesaurus, namely: text matching, ontology querying, and semantic visual querying. We evaluate the methods against the automatic search task of the TRECVID 2005 video retrieval benchmark, using a news video archive of 85 h in combination with a thesaurus of 363 machine learned concept detectors. We assess the influence of thesaurus size on video search performance, evaluate and compare the multimodal selection strategies for concept detectors, and finally discuss their combined potential using oracle fusion. The set of queries in the TRECVID 2005 corpus is too small for us to be definite in our conclusions, but the results suggest promising new lines of research.**

***Index Terms*—Concept learning, content analysis and indexing, knowledge modeling, multimedia information systems, video retrieval.**

I. Introduction

VIDEO has become the medium of choice in applications such as communication, education, and entertainment. In each of these, the video carries a semantic message which can be very versatile. For a human the meaning of the message is immediate, but for a computer that is far from true. This discrepancy is commonly referred to as the semantic gap [1].

Semantic video indexing is the process of automatically detecting the presence of a semantic concept in a video stream. It is impossible to develop a dedicated detector for each possible concept as there are just too many concepts. A recent trend in semantic video indexing has therefore been to search for generic methods that learn a detector from a set of examples [2]. This emphasis on generic indexing has opened up the possibility of moving to larger sets of concept detectors. MediaMill has published a collection of 101 machine-learned detectors [3]. LSCOM is working towards a set of 1000 detectors [4]. Both are learned from manually annotated examples from a news video corpus and have varying performance. Annotation constitutes a major effort and for any domain new concepts and new examples will have to be added. It is unrealistic to assume that such a purely data-driven approach will ever reach the richness of users' vocabularies.

This richness of vocabulary is also a well-known problem for humans describing video in words. A variety of terms are used to describe the same video fragment by different users, or by the same user in different contexts. Exploiting ontologies [5]–[7] to structure terms employed by users can make descriptions more consistent and can aid the user in selecting the right term for a semantic concept.

Our aim in this paper is to link a general-purpose ontology (with over 100 000 concepts) to a specific detector set (with several 100s of concepts). In this way, inherently uncertain detector results will be embedded in a semantically rich context. Hence, we can, for example, disambiguate various interpretations or find more general concepts. As the news domain is broad and can in theory contain any topic, a large and domain independent ontology is a must. As our ontology we use WordNet [5], a lexical database in which nouns, verbs, adjectives, and adverbs are organized into synonym sets (synsets) based on their meanings and use in natural language. We establish a link between WordNet and a set of 363 detectors learned from both MediaMill and LSCOM annotations.

The first to add semantics to detectors by establishing links with a general-purpose ontology were Hoogs *et al.* [8] who connected a limited set of visual attributes to WordNet. Combining low-level visual attributes with concepts in an ontology is difficult as there is a big gap between the two. In this paper we take a different, more intuitive approach: we link high-level concept detectors to concepts in an ontology. It should be noted, however, that detectors and the elements of an ontology are of a different nature. Detectors are uncertain whereas ontologies use symbolic facts. As a consequence they have been studied in completely different research fields. Having established a relation does not necessarily mean that the results of a task originating in one field will improve when augmented with techniques from the other field.

Our main research question therefore addresses the following: do semantically enriched detectors actually enhance results in semantic retrieval tasks? We evaluate retrieval results on 85 h of international broadcast news data from the 2005 TRECVID benchmark [9].

Manuscript received October 12, 2006; revised April 4, 2007. This work was supported by the BSIK MultimediaN Project, the NWO MuNCH Project, and the E.U. IST Programme of the 6th FP for RTD under Project MultiMATCH Contract IST-033104. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mohan S. Kankanhalli.

C. G. M. Snoek, B. Huurnink, M. de Rijke, and M. Worring are with the Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, 1098 SJ Amsterdam, The Netherlands (e-mail: cgmsnoek@science.uva.nl).

L. Hollink and G. Schreiber are with the Computer Science Department, FEW, Free University Amsterdam, 1081 HV Amsterdam, The Netherlands.

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.2007.900156

The paper is organized as follows. We discuss related work in Section II. We explain the process of adding semantics to detectors in Section III. We then present different strategies for selecting semantically enriched detectors for video retrieval in Section IV. Our experimental setup is presented in Section V and the experimental results in Section VI. We conclude in Section VII.

II. Related Work

Traditional video retrieval methods handle the notion of concepts implicitly. They extract low-level features from the video data and map this to a user query, assuming that the low-level features correspond to the high-level semantics of the query. Features can stem from textual resources that can be associated to video, like closed captions, or speech recognition results, e.g., [10], [11]. Alternatively, low-level visual features, e.g., color [12], texture [13], shape [14], and spatiotemporal features [15], are used in combination with query images. More recently, approaches have been proposed that combine text and image features for retrieval, e.g., [16]–[21]. We adhere to a multimedia approach also, but we use the notion of concepts explicitly, by expressing user queries in terms of high-level concept detectors rather than low-level features.

Such a high-level video retrieval approach requires detection of concepts. Early approaches aiming for concept detection focused on the feasibility of mapping low-level features, e.g., color, pitch, and term frequency, directly to high-level semantic concepts, like *commercials* [22], *nature* [23], and *baseball* [24]. This has yielded a variety of dedicated methods, which exploit simple decision rules to map low-level features to a single semantic concept. Generic approaches for concept detection [3], [25]–[29] have emerged as an adequate alternative for specific methods. Generic approaches learn a wide variety of concepts from a set of low-level features, which are often fused in various ways. In contrast to specific methods, these approaches exploit the observation that mapping multimedia features to concepts requires many decision rules. These rules are distilled using machine learning. The machine learning paradigm has proven to be quite successful in terms of generic detection [26], [28]. However, concept detection performance is still far from perfect; the state-of-the-art typically obtains reasonable precision, but low recall.

Learning requires labeled examples. To cope with the demand for labeled examples, Lin *et al.* initiated a collaborative annotation effort in the TRECVID 2003 benchmark [30]. Using tools from Christel *et al.* [31] and Volkmer *et al.* [32], [33] a common annotation effort was again made for the TRECVID 2005 benchmark, yielding a large and accurate set of labeled examples for 39 concepts taken from a predefined collection [4]. We provided an extension of this compilation, increasing the collection to 101 concept annotations, and also donated the low-level features, classifier models, and resulting concept detectors for this set of concepts on TRECVID 2005 and 2006 data as part of the MediaMill Challenge [3]. Recently, the LSCOM consortium finished a manual annotation effort for 1000 concepts [4]; concept detectors are expected to follow soon. This brings concept detection within reach of research in ontology engineering, i.e., creating and maintaining large, typically $10\ 000+$ structured sets of shared concepts.

Ontologies provide background knowledge about various topics. Examples are SnoMed, MeSH, the Gene Ontology and the metathesaurus UMLS for health care, AAT and Iconclass for art, and the generic ontologies WordNet and Cyc. Ontologies have various uses in the annotation and search process. Existing, well-established ontologies provide a shared vocabulary. The vocabulary terms and their meanings are agreed upon. Meaning is partially captured in the (hierarchical) structure of the ontology. Polysemous terms can be disambiguated, and relations between concepts in the ontology can be used to support the annotation and search process [34], [35]. Ontologies are currently being used for manual annotation [36], [37], and where manual annotations are not feasible or available, they have been used to aid retrieval based on captions or other text associated with the visual data [38]. These ontologies are, however, not suitable for semantic retrieval based on the visual properties of the data, since they contain little visual information about the concepts they describe.

Some work has been done to combine ontologies with visual features. Hoogs *et al.* [8] linked ontologies and visual features by manually extending WordNet with tags describing visibility, different aspects of motion, location inside or outside, and frequency of occurrence. In [39] a visual ontology was built that contains general and visual knowledge from two existing sources: WordNet and MPEG-7. Bertini *et al.* [40] propose a "pictorially enriched" ontology in which both linguistic terms and visual prototypes make up the nodes of the ontology. To the best of our knowledge, no work exists that links an ontology to the high-level concepts appearing in video data.

III. Adding Semantics to Detectors

Fig. 1 shows the schema used for semantically enriching concept detectors. We call the semantically enriched collection of concept detectors a *multimedia thesaurus*. It consists of textual descriptions, links to WordNet synsets, and visual models of the concept detectors, as detailed below.

A. Textual Descriptions

Each concept detector ω is associated with a manually created textual description, d_ω. It elaborates on the visual elements that should—or should not—be present. For example, the description for the concept detector *storms* is "outdoor scenes of stormy weather, thunderstorms, lightning." It explicitly indicates that video containing lightning and thunderstorms should be tagged as storms. The descriptions are by no means exhaustive, usually consisting of one or two sentences [3], [4], but they do contain a significant amount of information about the different kinds of visual content associated with each detector.

B. Links to Wordnet

We manually create links between concept detectors and WordNet synsets. To allow for scalability one prefers to obtain the link between concept detectors and WordNet synsets automatically. However, automatically mapping a concept detector to an ontology is still a difficult issue. The manual process guarantees high quality links, which are necessary to avoid

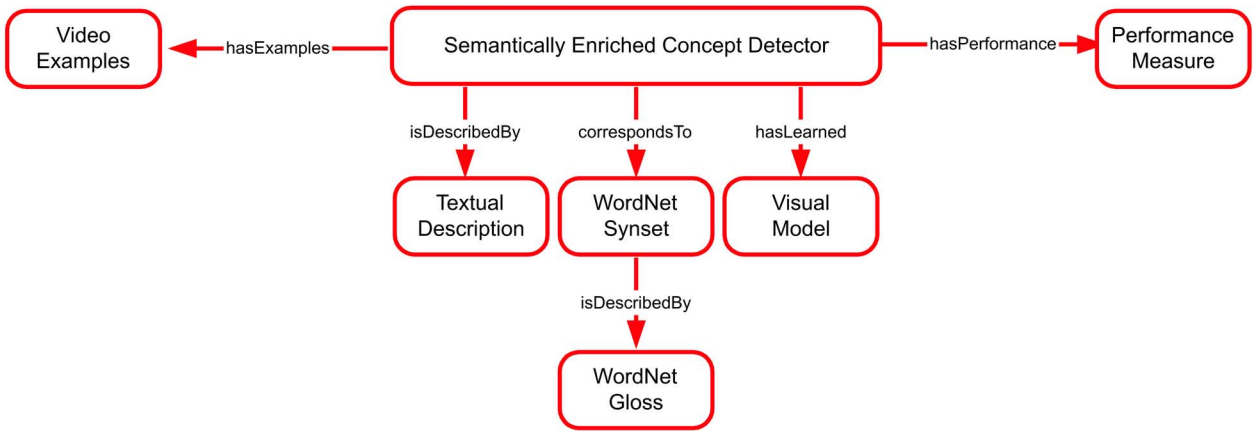

Fig. 1. Data model for semantically enriched detectors. A semantically enriched detector consists of a textual description, a link to WordNet, and a visual model. We refer to a collection of semantically enriched concept detectors as a multimedia thesaurus.

obscuring the experimental results. When automatic reasoning methods become available that automatically link concepts with high accuracy, these might at least partly substitute the manual process. The links, l_ω, are based on a comparison between the textual descriptions associated with each concept detector and WordNet "glosses," which are short descriptions of the synsets. Each concept is linked to 1–6 synsets, with at most two per part of speech (noun, verb, adjective). Concept detectors for specific persons that are not present in WordNet are linked as instances of a noun-synset. E.g., *Ariel Sharon* is not present in WordNet and is therefore linked as an instance of the noun-synset "Prime Minister." Each concept was linked to WordNet by two people independently. Overlap between the linkers was consistently around 65%, and the concepts without initial agreements were discussed until agreement was reached.

C. Visual Model

To arrive at a visual model v_ω for a concept detector, we build on previous work in generic concept detection, e.g., [3], [25]–[29]. Similar to this work, we view concept detection in video as a pattern recognition problem. Given a pattern $\vec{x}$, which is part of a shot, the aim is to obtain a confidence measure, $p(\omega|\vec{x})$, which indicates whether semantic concept ω is present in a shot.

Feature extraction is based on the method described in [3], [29], which is robust across different video data sets while maintaining competitive performance. We first extract a number of color invariant texture features per pixel. Based on these, we label a set of predefined regions in a key frame with similarity scores for a total of 15 low-level visual region concepts, resulting in a 15-bin histogram. We vary the size of the predefined regions to obtain a total of 8 concept occurrence histograms that characterize both global and local color-texture information. We concatenate the histograms to yield a 120-dimensional visual feature vector per key frame, $\vec{x}$.

For machine learning of concept detectors we adopt the experimental setup proposed in [3]. Hence, we divide a data set *a priori* into nonoverlapping training and validation sets. The training set $\mathcal{A}$ contains 70% of the data, and the validation set $\mathcal{B}$ holds the remaining 30%. We obtain the *a priori* concept occurrence by dividing the number of labeled video examples by the total number of shots in the archive. To obtain the confidence measure $p(\omega|\vec{x})$ we use the Support Vector Machine (SVM) framework [41]; see [3], [26], [28]. Here we use the LIBSVM implementation [42] with radial basis function and probabilistic output [43]. SVM classifiers thus trained for ω, result in an estimate $p(\omega|\vec{x},\vec{q})$, where $\vec{q}$ are parameters of the SVM. We obtain good parameter settings by performing an iterative search on a large number of SVM parameter combinations on training data. We measure performance of all parameter combinations and select the combination that yields the best performance after 3-fold cross validation. The result of the parameter search over $\vec{q}$ is the improved visual model $v_\omega = p(\omega|\vec{x},\vec{q}^*)$, contracted to $p^*(\omega|\vec{x})$.

Summarizing this section, a semantically enriched detector ω is defined as:

$$\omega = [d_\omega, l_\omega, v_\omega] \tag{1}$$

and the multimedia thesaurus Ω is the union over all ω.

IV. Detector Selection Strategies

In the video retrieval paradigm, user queries may consist of example videos, natural language text, or both. Although current practice suggests that combining concept detectors with traditional text and image retrieval techniques [44], [45] may yield improved performance, they might equally well hurt performance as none of these techniques is perfect yet. Speech recognition for the Dutch language, for example, is still problematic. We therefore opt for automatic selection of a concept detector appropriate to the query, allowing users to quickly retrieve a list of relevant video fragments. We focus on the selection of a single best detector to maximize retrieval performance, and base our selection methods on the modalities associated with the user query: the textual modality and the visual modality. We also try to model the original user intent motivating the query by using ontology knowledge.

Based on the different query modalities and the user intent we identify three different approaches for selecting the most appropriate detector, as shown in Fig. 2. In the textual modality we use

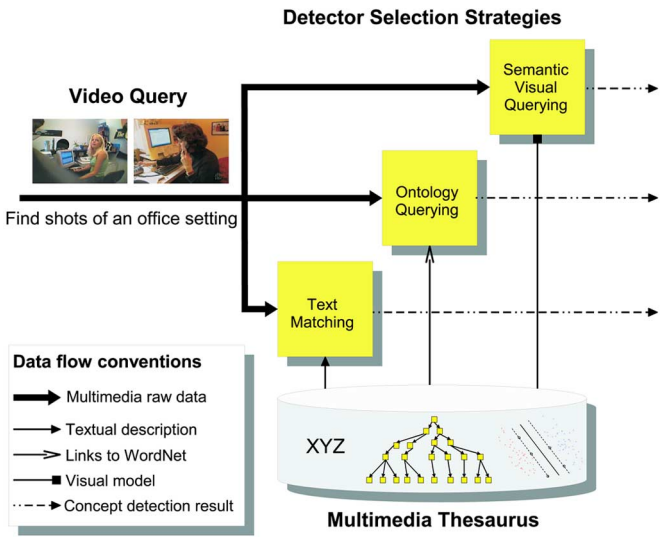

Fig. 2. Three different strategies for selecting a semantically enriched concept detector from a multimedia thesaurus, given a multimodal user query: text matching, ontology querying, and semantic visual querying.

a detector selection method based on text matching. When modeling the user's intent, we elicit semantic knowledge through natural language analysis and ontology linking, using this to create a detector selection method based on ontology querying. In the visual modality we use a detector selection method based on semantic visual queries. Below, we detail our detector selection strategies.

A. Selection by Text Matching

As in our multimedia thesaurus each detector is associated with a textual description d_ω, we can match the text specification of a query with the textual description of a detector. Both the description and the query text are normalized: commonly occurring words are removed using the SMART stop list [46], all text is converted to lower case, and punctuation is removed. Each detector description, or document, is represented by a term vector, where the elements in the vector correspond to unique normalized words, or terms. The concept descriptions are written in natural language—as such, the term distribution roughly corresponds with Zipf's law. Therefore, the vector space model [47], which discounts for frequently occurring terms and emphasizes rare ones, is appropriate to match the words in the user query to words in the detector descriptions.

Specifically, with a collection of descriptions D, a candidate description d_ω in D and query q containing terms t_i, we use the following implementation of the vector space model [48]:

$$\mathrm{sim}(q, d_\omega) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{\mathrm{norm}_q} \cdot \frac{tf_{t,d_\omega} \cdot idf_t}{\mathrm{norm}_d} \cdot \mathrm{coord}_{q,d} \quad (2)$$

where

$$tf_{t,X} = \sqrt{\mathrm{freq}(t, X)} \quad \mathrm{norm}_d = \sqrt{|d_\omega|}$$
$$idf_t = 1 + \log \frac{|D|}{\mathrm{freq}(t, D)} \quad \mathrm{coord}_{q,d} = \frac{|q \cap d_\omega|}{|q|}$$
$$\mathrm{norm}_q = \sqrt{\sum_{t \in q} tf_{t,q} \cdot idf_t{}^2}.$$

We select the detector with the highest similarity between the query vector and the description vector, $\mathrm{sim}(q, d_\omega)$, from multimedia thesaurus Ω:

$$\omega_d = \arg\max_{\omega \in \Omega} \quad \mathrm{sim}(q, d_\omega). \quad (3)$$

B. Selection by Ontology Querying

When designing a detector selection method based on ontology querying, we attempt to model the user intent from the query. We first perform syntactic disambiguation of the words in the text query. The memory-based shallow parser described in [49] is used to extract nouns and noun chunks from the text. These are then translated to ontological concepts. First, we look up each noun in WordNet. When a match has been found the matched words are eliminated from further lookups. Then, we look up any remaining nouns in WordNet. The result is a number of WordNet noun-synsets related to the query text.

As described in Section III-B, the concept detectors are also linked to WordNet synsets. We now query the ontology to determine which concept detector is most related to the original query text.[1] Here, we must define what "most related" means. Simply counting the number of relations between a query-synset and a concept-detector-synset does not give a good indication of relatedness, since the distances of the relations in WordNet are not uniform. In addition, we encounter the problem of distinguishing between concept detectors that are equally close to the textual query. To overcome this we use Resnik's measure of information content [50], where a concept is viewed as the composite of its synonyms and its sub-concepts. E.g., *vehicle* is defined not only by all occurrences of the word "vehicle", but also by all occurrences of the words "car," "truck," "SUV," and so on. The information content of a concept is negative the log likelihood of that concept occurring in a tagged text, where the likelihood of a concept is defined in terms of occurrences of that concept and all subconcepts, or subsumers, of that concept:

$$\mathrm{p}(l_\omega) = \frac{\sum_{n \in \mathrm{words}(l_\omega)} \mathrm{count}(n)}{N} \quad (4)$$

where l_ω is a linked concept, $\mathrm{words}(l_\omega)$ is the set of all noun lemmas belonging to l_ω and all subsumers of l_ω, N is the total number noun lemmas n observed in an external corpus, and $\mathrm{count}(n)$ is the number of times each member of $\mathrm{words}(l_\omega)$ is observed in the external corpus. We used the SemCor news corpus [51] as our external corpus. We select the concept detector that maximizes information content:

$$\omega_l = \arg\max_{\omega \in \Omega} \quad (-\log \mathrm{p}(l_\omega)). \quad (5)$$

C. Selection by Semantic Visual Querying

Concept detectors may also be selected by using semantic visual querying. Although it is hard to expect that general users will prefer to provide a number of image examples rather than explicitly specifying the semantic concept they need, semantic visual querying might prove a valuable additional strategy when other selection strategies fail. For semantic visual querying we

[1]An RDF/OWL representation of the ontology can be queried at http://www.cs.vu.nl/~laurah/semantics2detectors.html/

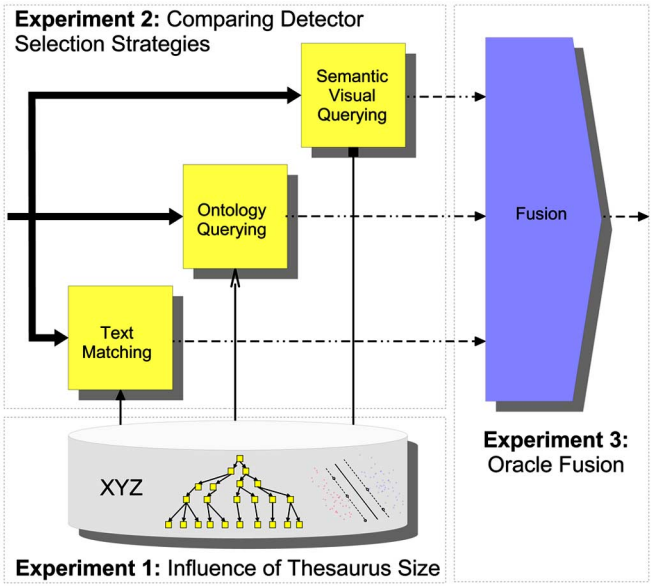

Fig. 3. Schematic overview of our video retrieval experiments, using the conventions of Fig. 2. In experiment 1 we assess the influence of an increasing thesaurus size on video retrieval performance. In experiment 2 we evaluate three concept detector selection strategies. Experiment 3 explores an oracle fusion of individual detector selection methods.

follow the approach by Rasiwasia *et al.* [52]. In this scenario all available visual models are applied to the query image; next, the model with the highest posterior probability is selected as most relevant. In our implementation, concept detector selection based on semantic visual querying first extracts visual features from the query images $\vec{f}$, as explained in Section III-C. Based on the features, we predict a posterior concept probability for each query image. We select the detector with the maximum posterior probability:

$$\omega_v = \arg\max_{\omega \in \Omega} \quad p^*(v_\omega|\vec{f}). \tag{6}$$

V. Experimental Setup

For evaluation we use the automatic search task of the 2005 TREC Video Retrieval Evaluation (TRECVID) [9]. Rather than aiming for the best possible retrieval result, our goal is to assess the influence of adding semantics to detectors. To that end, our experiments focus on the evaluation of strategies for selection of a single concept detector, given an information need. We first determine the best possible single concept detector for an information need, or topic, given an increasing thesaurus of concept detectors. Then, we assess different algorithms for the three strategies described in Section IV and select the best implementation for each strategy. We compare the individual approaches; analyzing their strengths and weaknesses. Finally, we explore a combination method that fuses individual detector results. A schematic overview of the experiments is depicted in Fig. 3. We will now detail the search task, data set, multimedia thesaurus, and our experiments.

A. TRECVID Automatic Video Search Task

The goal of the search task is to satisfy a number of video information needs. Given such a need as input, a video search engine should produce a ranked list of results without human intervention. The 2005 search task contains 24 search topics in total. For each topic we return a ranked list of up to 1000 results. The ground truth for all 24 topics is made available by the TRECVID organizers, and to assess our retrieval methods we use *average precision* (AP), following the standard in TRECVID evaluations [9]. 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 results. Hence, it combines precision and recall into one performance value. We report the mean average precision (MAP) over all search topics as an indicator for overall search system performance.

B. Data Set & Multimedia Thesaurus Building

The TRECVID 2005 video archive contains 169 h of video data, with 287 episodes from 13 broadcast news shows from US, Arabic, and Chinese sources, recorded during November 2004. The test data collection contains approximately 85 h of video data. The video archives come accompanied by a common camera shot segmentation, which serves as the unit for retrieval. We face the task of specifying a set of semantic concept detectors for the TRECVID 2005 data set. We adopt the set of 101 concept detectors made publicly available as part of the MediaMill Challenge [3]. These use the implementation sketched in Section III-C. Using the same method, we learn concept detectors based on the manual annotations of LSCOM [4]. Concept detectors in both sets of annotations are related to program categories, settings, people, objects, activities, events, and graphics. Concepts are added to the combined thesaurus only when at least 30 positive instances are identified in the TRECVID 2005 training set. When concepts in the MediaMill and LSCOM thesauri link to the same WordNet synset they are considered to be similar. In those cases, the performance on validation set $\mathcal{B}$ is used as selection criterion. This process results in a combined thesaurus of 363 concept detectors.

C. Experiments

We investigate the impact of adding semantics to detectors by performing the following three experiments.

- **Experiment 1:** *What is the Influence of Increasing the Concept Detector Thesaurus Size for Video Search?*

To assess the influence of growing concept detector thesauri on video retrieval performance we randomly select a bag of 10 concepts from our thesaurus of 363 detectors. We evaluate each detector in the bag against all 24 search topics and determine the one that maximizes AP for each topic. Hence, we determine the upper limit in MAP score obtainable with this bag. In the next iteration, we select a random bag of 20 concept detectors from the thesaurus, and once more the optimal MAP is computed. This process is iterated until all concept detectors have been selected. To reduce the influence of random effects, which may disturb our judgement of increasing thesaurus size on video search performance, we repeat the random selection process 100 times.

- **Experiment 2:** *How to Select the Most Appropriate Concept Detector for a Video Search Query?*

For each of the three modalities identified in Fig. 2, we want to identify the most appropriate concept detector. Hence, our

second experiment consists of three sub-experiments, as detailed below, and a fourth sub-experiment that compares the three individual methods.

1) Experiment 2a: What Is the Most Appropriate Concept Detector Using Text Matching?: We assess the influence of text matching on concept detector selection by indexing the concept detector descriptions in the Lucene [48] search engine, using the implementation described in Section IV-A. Within text retrieval, collections are generally quite large compared to the 363 concept descriptions that we have available. We hypothesize that in this small collection, where there are comparatively few text terms to match, recall is a bigger issue than in large collections. Effective ways to increase recall are stemming, where words are reduced to their root forms, and character n-gramming, where words are iteratively broken up into sequences of n characters. We perform three experiments for text matching—perfect match, stemmed match, and character n-gram match. For stemming we use the Porter stemming algorithm [53]. For character n-grams we use sequences of four characters as this approach has been shown to perform well for English [54].

2) Experiment 2b: What Is the Most Appropriate Concept Detector Using Ontology Querying?: As described in Section IV-B, we query the ontology for the concept detector most closely related to the noun-synsets in the query. Several approaches exist for estimating the semantic distance between synsets (see for example [55]). In this paper, we employ two approaches that have shown promising results in earlier studies. The first uses Resnik similarity, which is a measure of semantic similarity in an is-a taxonomy based on information content [50] (see Section IV-B); in the case of WordNet, the "is-a taxonomy" translates to the hyponym/hypernym hierarchy. The second approach uses subsumption relations (hyponym/hypernym) as well as part-of relations. While the use of hyponym relations is commonly accepted, a recent study [35] showed that the inclusion of part-of and hypernym relations further improves retrieval results, especially for visual data. A concept detector directly matching the query synset is considered closest. After that, a concept detector that has a hypernym relation to the query synset is considered closest, followed by a concept detector that has a hyponym or part-of relation to the query synset. Many queries consist of more than one noun synset. When this is the case, we first seek the closest concept detector that is related to the first query synset. If there are no matching concept detectors, we proceed to the next query synset, until a detector is found or the last synset has been reached. In addition, we test two methods to break ties between detectors that are equally close to a query synset: 1) the information content of the concept detector and 2) the *a priori* chance that a concept is present in our data set.

3) Experiment 2c: What Is the Most Appropriate Concept Detector Using Semantic Visual Querying?: Selecting concept detectors using semantic visual querying may be a brittle approach when concepts are not distributed equally in the data set, as is often the case in realistic video retrieval applications. Rather than selecting the concept with the maximum score—which is often the most robust but also the least informative one, e.g., *person, face, outdoor*—we also assess a heuristic selection mechanism that takes concept frequency into account. Similar to the vector space model used in Section IV-A, we discount for frequently occurring terms and we emphasize rare ones. We take the posterior probability as a substitute for term frequency and divide by the logarithm of the inverse concept frequency. By doing so, we prioritize less frequent, but more discriminative, concepts with reasonable posterior probability scores over frequent, but less discriminative, concepts with high posterior probability scores.

4) Experiment 2d: What Are the Strengths and Weaknesses of the Selection Strategies?: We compare the three different selection strategies quantitatively as well as qualitatively. Based on previous TRECVID search results [9], [16]–[21], we anticipate that the AP varies highly per topic. Therefore, we normalize the AP scores of the three methods by dividing them by the AP score of the best possible detector. These percentages give a better indication of the differences between the methods than the raw AP data. This has the added advantage that unreliable statistical results due to outliers are avoided.

We examine whether there are significant differences between the three detector selection methods. Since the data are not normally distributed we perform a nonparametric Kruskal-Wallis test. We also perform pairwise Wilcoxon signed rank tests. We look for correlation between the three selection methods with Spearman's rank correlation coefficient. Finally, we qualitatively examine the differences by looking at which detectors are selected by the three methods.

- **Experiment 3:** *What is the Influence of Combining Detector Selection Strategies?*

Since the individual concept detector selection strategies in experiment 2 work with different modalities, it is natural to ask to which extent they complement each other. A combination of some or all of them could further improve video retrieval performance [56]. Various combination methods exists; the linear combination of individual methods is often evaluated as one of the most effective combination methods, see for example [57], [58]. We adopt a linear combination function, similar to [21], [58], which uses a single combination factor λ_1 for pair-wise combination of two concept detectors, defined as:

$$p_2^*(\omega_1, \omega_2|\vec{x}) = \lambda_1 \cdot p^*(\omega_1|\vec{x}) + (1 - \lambda_1) \cdot p^*(\omega_2|\vec{x}) \quad (7)$$

where $\lambda_1 \in [0, 1]$. To allow for three-way combination of selected concept detectors we extend (7) with an additional combination factor λ_2, defined as:

$$p_3^*(\omega_1, \omega_2, \omega_3|\vec{x}) = \lambda_1 \cdot p^*(\omega_1|\vec{x}) + \lambda_2 \cdot p^*(\omega_2|\vec{x}) + (1 - (\lambda_1 + \lambda_2)) \cdot p^*(\omega_3|\vec{x}) \quad (8)$$

where $\lambda_2 \in [0, 1]$, and $\lambda_1 + \lambda_2 \leq 1$. To assess the influence of combining detector selection mechanisms, we perform an experiment that evaluates all possible linear combinations with steps of 0.1 for both λ_1 and λ_2. We term this combination "oracle fusion" as it uses the test set results to select the optimal combination on a per-query basis. It is included to explore the upper limits of performance that are reachable by combining detector selection strategies.

We compare the oracle fusion experiments using a Kruskal-Wallis test and pairwise Wilcoxon signed rank tests. Wilcoxon's test is also used to examine differences between

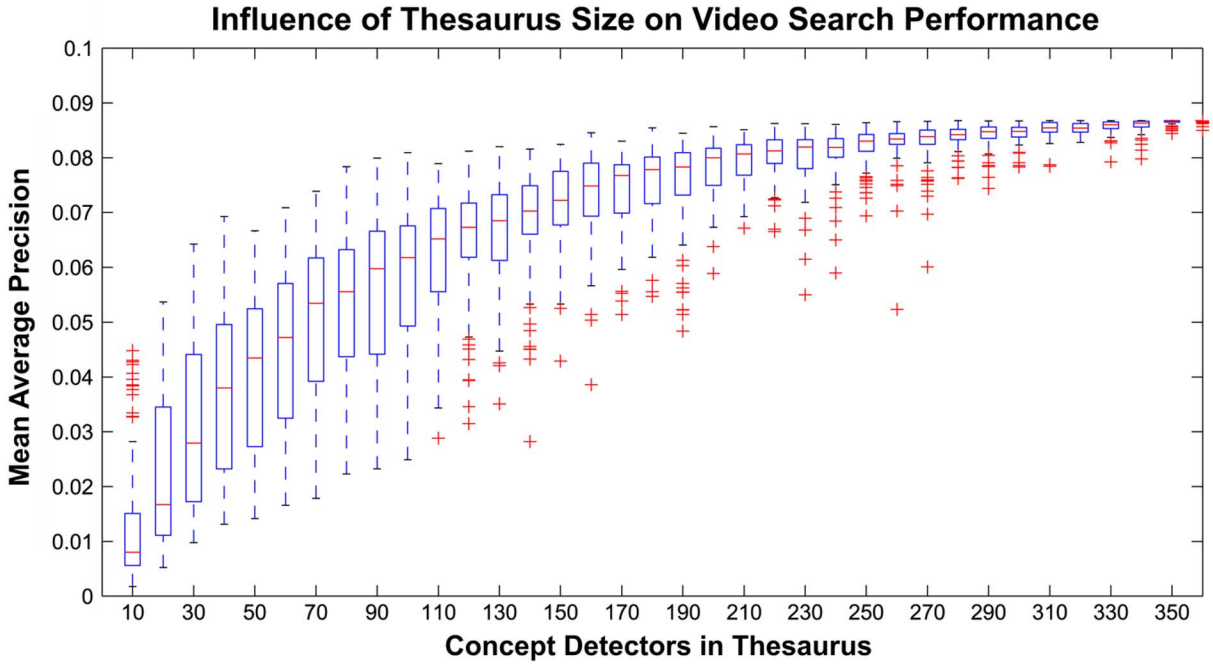

Fig. 4. Box plot showing the positive influence of an increasing thesaurus size, in random bags of 10 machine learned concept detectors, on MAP over 24 topics from the TRECVID 2005 video retrieval benchmark. Extreme values after 100 repetitions are marked (+) as outliers.

the results of the fusion experiments and the results of the single-method experiments.

VI. Results

A. Experiment 1: What Is the Influence of Increasing Concept Detector Thesaurus Size for Video Search?

We summarize the influence of an increasing thesaurus of concept detectors on video search performance in the box plot in Fig. 4. There is a clear positive correlation between the number of concept detectors in the thesaurus and video retrieval performance. The box plot also shows that the median is shifted towards the bottom of the box for the first 30 concept detectors, even when the outliers are ignored. This indicates that, on average, performance is low for small thesauri, but some detectors perform well for specific topics. However, it is unlikely that a large variety of topics can be addressed with a small thesaurus, which explains the skew. With only 10 randomly selected concept detectors the median MAP score is 0.008. Indeed, the usage of few concept detectors is of limited use for video retrieval. However, a steady increase in thesaurus size has a positive influence on search performance. For the first 60 concept detectors this relation is even linear, increasing MAP from 0.008 to 0.047. When thesauri grow, more search topics can be addressed with good performance. However, the shift towards the high end of the box indicates that a substantial number of concept detectors in our thesaurus do not perform accurate enough, yet, to be decisive for performance. As a result, when more than 70 concept detectors are added, the increase is less strong, but it keeps rising until the limit of this thesaurus is reached for the maximum obtainable MAP of 0.087. Note that this value is competitive with the state-of-the-art in video search [9].

B. Experiment 2: How to Select the Most Appropriate Concept Detector for a Video Search Query?

Due to lack of space we are not able to provide detailed breakdowns of scores for all our experiments. Table I lists the AP scores for the selected concept detector methods (columns 3–5) and for the best possible single detector (column 2).

1) Experiment 2a: What Is the Most Appropriate Concept Detector Using Text Matching?: Contrary to our expectations, we found that using exact text matching provided the best results with a MAP score of 0.0449, versus 0.0161 for stemmed text, and 0.0290 for n-grammed text. It appears that when retrieving detector descriptions using query text, it is more important to get exact matches to the original query terms than it is to aim for recall and increase the number of detector descriptions retrieved. We expect that this is due to our choice to select only a single best concept detector match. If we allow multiple detectors to be returned, techniques such as stemming and n-gramming might have a beneficial impact.

In the remainder we will use the exact text matching approach for concept detector selection using textual matches.

2) Experiment 2b: What Is the Most Appropriate Concept Detector Using Ontology Querying?: The approach using Resnik similarity (approach 1) was outperformed by the approach using subsumption/part-of relations (approach 2) regarding mean average precision (0.0218 and 0.0485, respectively), but the difference was not statistically significant. Examining the selected detectors, we see that approach 1 performs better on person x queries, while approach 2 benefits from the use of hypernyms.

A comparison between the use of information content to the use of *a priori* chances for distinguishing between concept detectors that are equally close to the topic, shows that the differences are minimal. Only four topics get a different detector,

TABLE I
COMPARISON OF THREE DETECTOR SELECTION STRATEGIES FOR VIDEO RETRIEVAL. SEARCH RESULTS ARE COMPARED AGAINST THE BEST POSSIBLE CONCEPT DETECTOR SCORE FOR EACH TOPIC IN RELATIVE PERCENTAGES OF AVERAGE PRECISION (AP%). THE BEST RESULT IS GIVEN IN BOLD

			Detector Selection Strategies					
	Best Possible		2a: Text Matching		2b: Ontology Querying		2c: Semantic Visual Querying	
Search Topic	*Best Detector*	*AP*	*Selected Detector*	*AP%*	*Selected Detector*	*AP%*	*Selected Detector*	*AP%*
Two visible tennis players on the court	Athlete	0.6501	Tennis Game	89.7%	**Athlete**	**100.0%**	Tennis Game	89.7%
A goal being made in a soccer match	Stadium	0.3429	Soccer Game	31.7%	Soccer Game	31.7%	**Grass**	**51.0%**
Basketball players on the court	Indoor Sports Venue	0.2801	Court	0.0%	Athlete	30.4%	**Basketball Game**	**81.5%**
A meeting with a large table and people	Furniture	0.1045	**Conference Room**	**73.6%**	Meeting	24.8%	Flag	1.0%
People with banners or signs	People Marching	0.1013	**Demonstration or Protest**	**73.7%**	Group	5.3%	Desert	0.4%
One or more military vehicles	Armored Vehicles	0.0892	**Tanks**	**38.1%**	**Tanks**	**38.1%**	Charts	0.0%
Helicopter in flight	Helicopters	0.0791	Helicopter Hovering	53.1%	**Helicopters**	**100.0%**	Helicopter Hovering	53.1%
A road with one or more cars	Car	0.0728	Car Crash	7.9%	**Road**	**65.9%**	Helicopters	4.4%
An airplane taking off	Classroom	0.0526	Airplane Flying	10.8%	Airplane Flying	10.8%	**Helicopters**	**87.3%**
A tall building	Office Building	0.0469	Tower	89.8%	**Building**	**98.8%**	Grass	0.2%
A ship or boat	Cloud	0.0427	**Boat or Ship**	**46.5%**	**Boat or Ship**	**46.5%**	Cigar Boats	39.5%
George Bush entering or leaving vehicle	Rocket Propelled Grenades	0.0365	**George Bush jr**	**6.6%**	**George Bush jr**	**6.6%**	Helicopter Hovering	0.0%
Omar Karami	Chair	0.0277	Ariel Sharon	0.8%	Ariel Sharon	0.8%	**Yasser Arafat**	**3.5%**
Graphic map of Iraq, Baghdad marked	Graphical Map	0.0269	**Graphical Map**	**100.0%**	**Graphical Map**	**100.0%**	**Graphical Map**	**100.0%**
Condoleeza Rice	US National Flag	0.0237	-	0.0%	-	0.0%	**Capitol**	**0.4%**
One or more palm trees	Weapons	0.0225	Tropical Setting	1.6%	Trees	23.4%	**Fire Weapon**	**44.3%**
Something on fire with flames and smoke	Violence	0.0151	**Smoke**	**95.1%**	Vehicle	41.4%	Soccer Game	18.9%
Mahmoud Abbas	Conference Room	0.0134	Ariel Sharon	0.5%	Ariel Sharon	0.5%	**Yasser Arafat**	**2.3%**
Hu Jintao	Iyad Allawi	0.0123	Hu Jintao	4.3%	George Bush sr	2.4%	**Non-US National Flags**	**55.0%**
People shaking hands	Beards	0.0110	Handshaking	14.6%	Group	10.2%	**Yasser Arafat**	**18.0%**
Office setting	Computers	0.0095	**Computers**	**100.0%**	Office	90.4%	Emile Lahoud	1.9%
Iyad Allawi	Iyad Allawi	0.0095	**Iyad Allawi**	**100.0%**	Ariel Sharon	46.6%	**Iyad Allawi**	**100.0%**
Tony Blair	Election Campaign Address	0.0067	Tony Blair	0.0%	Tony Blair	0.0%	**George Bush jr**	**29.6%**
People entering or leaving a building	Muslims	0.0044	USA Government Building	6.4%	**Group**	**27.0%**	Reporters	8.5%
Mean		*0.0867*		*50.8%*		***56.0%***		*55.6%*
Number of highest scores		.		*9*		*9*		***12***

and the difference in MAP is only 0.0034. A possible explanation is that for most topics we find one concept detector that is closest to the topic synsets, which means that neither information content, nor *a priori* chances have to be used. In the remaining sections, we continue with the results of the subsumption/part-of approach using information content, since this gives us the highest AP scores.

Using this approach, a detector was found for all but one of the queries of TRECVID 2005 that is at most one hypenym/hyponym/part-of relation away from a topic synset. This suggests that our large detector pool has a good coverage of the TRECVID queries.

3) Experiment 2c: What Is the Most Appropriate Concept Detector Using Semantic Visual Querying?: We observe that selection of concept detectors from semantic visual examples profits from a normalization step that takes *a priori* concept occurrence into account. When we do not normalize the posterior probability, selection based on semantic examples picks in 23 out of 24 queries (data not shown) one of the four most frequent concepts appearing in this data set, namely *people, face, overlayed text*, or *outdoor* [3]. While this is often correct, the concept is so general that it hardly contributes to retrieval performance. The only exception is the search topic for tennis players, where the selected *sport games* detector has good AP.

When we take *a priori* concept frequency into account, search results improve. Results of this experiment are summarized in the last column of Table I. We observe that selected detectors sometimes accurately reflect the semantics of the search topics, e.g., *Iyad Allawi, Graphical Map, Tennis Game, Helicopter Hovering, Cigar Boats, Basketball Game*, and *Grass*. This is not always the case however, and questionable detectors are selected for some search topics. This especially hurts the person x queries; for the topic *find shots of George Bush entering or leaving a vehicle*, for example, the optimal detector is *rocket propelled grenades*. However, a detector that matches well in terms of semantics is no guarantee for good search performance. In cases such as *find shots of graphical maps with Baghdad marked* or *find shots of ships*, the selected

TABLE II
COMPARISON OF PAIR-WISE (7) AND THREE-WAY (8) ORACLE FUSION OF THE DETECTOR SELECTION STRATEGIES FROM TABLE I. SEARCH RESULTS ARE COMPARED, WITH VARYING λ_1 AND λ_2, AGAINST THE BEST POSSIBLE CONCEPT DETECTOR SCORE FOR EACH TOPIC IN RELATIVE PERCENTAGES OF AVERAGE PRECISION (AP%). FUSION RESULTS THAT RELY ON ONE DETECTOR ONLY ARE INDICATED WITH—. THE BEST RESULT IS GIVEN IN BOLD

			Oracle Fusion of Detector Selection Strategies								
	Best Possible		**2a + 2b**		**2a + 2c**		**2b + 2c**		**2a + 2b + 2c**		
Search Topic	*Best Detector*	*AP*	λ_1	*AP%*	λ_1	*AP%*	λ_1	*AP%*	λ_1	λ_2	*AP%*
Two visible tennis players on the court	Athlete	0.6501	0.7	**105.4%**	—	89.7%	0.3	**105.4%**	0.0	0.3	**105.4%**
A goal being made in a soccer match	Stadium	0.3429	—	31.7%	0.3	**76.5%**	0.3	**76.5%**	0.0	0.3	**76.5%**
Basketball players on the court	Indoor Sports Venue	0.2801	0.9	30.4%	0.2	81.6%	0.2	**86.1%**	0.0	0.2	**86.1%**
A meeting with a large table and people	Furniture	0.1045	—	**73.6%**	—	**73.6%**	0.9	25.0%	—	—	**73.6%**
People with banners or signs	People Marching	0.1013	—	**73.7%**	—	**73.7%**	0.6	5.3%	—	—	**73.7%**
One or more military vehicles	Armored Vehicles	0.0892	—	**38.1%**	—	**38.1%**	—	**38.1%**	—	—	**38.1%**
Helicopter in flight	Helicopters	0.0791	—	**100.0%**	—	53.1%	—	**100.0%**	—	—	**100.0%**
A road with one or more cars	Car	0.0728	0.9	**66.9%**	—	7.9%	0.5	66.6%	0.9	0.1	**66.9%**
An airplane taking off	Classroom	0.0526	—	10.8%	—	**87.3%**	—	**87.3%**	—	—	**87.3%**
A tall building	Office Building	0.0469	0.8	**141.2%**	—	89.8%	0.9	98.8%	0.8	0.2	**141.2%**
A ship or boat	Cloud	0.0427	—	46.5%	0.1	**55.8%**	0.1	**55.8%**	0.0	0.1	**55.8%**
George Bush entering or leaving vehicle	Rocket Propelled Grenades	0.0365	—	**6.6%**	0.6	**6.6%**	0.6	**6.6%**	0.0	0.6	**6.6%**
Omar Karami	Chair	0.0277	—	0.8%	0.9	**4.0%**	0.9	**4.0%**	0.0	0.9	**4.0%**
Graphic map of Iraq, Baghdad marked	Graphical Map	0.0269	—	**100.0%**	—	**100.0%**	—	**100.0%**	—	—	**100.0%**
Condoleeza Rice	US National Flag	0.0237			—	**0.4%**	—	**0.4%**	—	—	**0.4%**
One or more palm trees	Weapons	0.0225	0.1	23.4%	0.9	48.7%	0.8	49.7%	0.5	0.4	**53.2%**
Something on fire with flames and smoke	Violence	0.0151	0.9	100.7%	0.9	102.6%	0.7	38.4%	0.8	0.1	**103.0%**
Mahmoud Abbas	Conference Room	0.0134	—	0.5%	0.9	**2.4%**	0.9	**2.4%**	0.0	0.9	**2.4%**
Hu Jintao	Iyad Allawi	0.0123	0.9	5.5%	0.9	55.5%	0.8	55.5%	0.4	0.4	**56.2%**
People shaking hands	Beards	0.0110	—	14.6%	0.9	19.6%	0.1	**29.7%**	0.0	0.1	**29.7%**
Office setting	Computers	0.0095	0.1	**154.9%**	—	100.0%	—	90.4%	0.1	0.9	**154.9%**
Iyad Allawi	Iyad Allawi	0.0095	0.1	121.1%	—	100.0%	0.9	**121.1%**	0.0	0.9	**121.1%**
Tony Blair	Election Campaign Address	0.0067	—	0.0%	0.9	**29.7%**	0.9	**29.7%**	0.0	0.9	**29.7%**
People entering or leaving a building	Muslims	0.0044	0.9	28.2%	0.4	9.2%	0.6	28.3%	0.8	0.1	**29.3%**
Mean		*0.0867*		*65.5%*		*72.4%*		*75.9%*			***83.4%***
Number of highest scores		.		*10*		*12*		*15*			*24*

detectors fit the topic, but perform only moderately well. In the first case the detector is not specific enough, in the second case its performance is not good enough. These results suggest that a measure is needed indicating when incorrect optimal detectors should be preferred over correct ones with bad video search results.

4) Experiment 2d: What Are the Strengths and Weaknesses of the Selection Strategies?: We found no significant differences between the results of the three individual selection experiments. Experiments 2a-2b, 2a-2c and 2b-2c also failed to show differences. We found a positive correlation between experiments 2a-2b, which was lacking between 2a-2c and 2b-2c. This suggests that the text-based and WordNet-based concept detector selection methods perform well on the same set of topics (and perform badly on the same set of topics) while the visual method scores well on other topics. This is supported by the fact that the text-based and WordNet-based methods select the same detector for 10 topics, while the visual method agreed on a detector only four times with the text-based method and only once with the WordNet based method.

C. Experiment 3: What Is the Influence of Combining Detector Selection Strategies?

We summarize the results of our combination experiments in Table II. The increase in MAP for all fusion experiments indicates that combining detector selection strategies pays off in

general. Pair-wise combination is especially effective when two different concept detectors obtain good average precision in isolation. For search topics such as *find shots with tall buildings* and *find shots of an office setting* the pair-wise combination of detectors selected by text matching and ontology querying even improves substantially upon the best possible single detector. A combination of selection by ontology querying and selection using semantic visual examples yields the most effective pair-wise combination strategy in terms of overall performance. However, no significant differences were found between the three types of pair-wise combination results. Since a large overlap in selected detectors exists between the three different selection strategies, three-way combination often boils down to pair-wise combination. For those search topics where three different concept detectors are selected, e.g., *find shots of palm trees*, three-way combination yields a further, but modest, increase over the best pair-wise combination. Again, no significant difference was found between pair-wise and three-way combination. However, using a Wilcoxons signed rank test, we did find significant differences between the results of the combination experiments and the results of the single-method experiments. The fusion experiments were consistently better at the 0.01 α-level.

VII. Discussion and Conclusion

We view this paper as a first step in a novel multidisciplinary approach to tackle the problem of semantic video retrieval. The results are not conclusive in the sense that they provide a solid basis for preferring a particular approach over others.

Experiment 1 gives indications about the number of thesaurus concepts (= thesaurus size) needed for optimal video retrieval performance. In Fig. 4 we can see that a thesaurus size of 100–200 already comes close to maximum performance. However, our experiments consider only 24 topics. A wider range of topics will likely require a larger thesaurus size to reach this same performance level.

In the detector selection experiment, experiment 2 we see that both in terms of MAP and in terms of the highest number of "best detector selections" the three selection strategies show comparable results. For text matching (2a) we found that exact matching works best, but this is probably a consequence of the fact that we select only a single detector. For ontology querying (2b) it is interesting to note the distinction between the hyponym/part-of and the Resnik method; the former performing best on "general concept" queries, the latter on "person x" queries. This suggests the use of a more refined concept-detector selection mechanism. Semantic visual querying (2c) was shown to correlate better with a different set of topics than both text matching and ontology querying. For this selection method we note the importance of avoiding frequently occurring but nondiscriminative concept detectors, such as for *people* and *outdoor*.

The fusion experiments (3) clearly show that we can gain by combining selection methods. It indicates that we can come close to achieving optimal concept-detector selection scores if we understand the situations in which it is useful to combine selection mechanisms. We should consider including a "selector-of-selector" step based on the query topic, which would propose a (combination of) selection method(s) that is likely to be optimal. At the moment, the set of topics included in this study provides insufficient information as a basis for such a meta-selection. More experimentation will be needed to clarify in which cases (e.g., for which classes of topics) two or more selection methods benefit from combination. The goal should be to identify, for example, whether topics involving named people require a different selection method than a topic involving a general concept such as *road*. Studying the nature of query topics might also reveal whether we are missing out important other categories of topics.

One limitation of our approach is that we have only considered situations in which the three individual methods select precisely one detector. This is likely to have been too strong. It is easy to imagine situations in which the selection strategies produce a set of multiple detectors. In principle, this would make it possible to get a higher average precision score than that of a single detector (which is the maximum score we can achieve in this study). However, a major increase in detection performance is needed before concept detector combination is really successful. We are planning experiments in which we lift this limitation.

Adopting a broader perspective, we also need to consider the effect of the domain we are working in. News video is a domain with special characteristics. The stylized shots, the highly domain-specific concepts (*female anchor*) and other factors are likely to make it difficult to predict how our methods would behave in other video retrieval domains, such as documentaries.

Finally, coming back to the research question we started with: have we shown that semantically-enriched detectors enhance results in semantic retrieval tasks? Our results do not yet permit us to respond with a firm "yes" to this question, but the results are encouraging. We have scratched the surface of a semantic video retrieval approach which combines different techniques. The results suggest promising new lines of research.

References

[1] A. W. M. 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. Machine Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] M. Worring and G. Schreiber, "Semantic image and video indexing in broad domains," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 909–910, Aug. 2007.

[3] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proc. ACM Multimedia*, Santa Barbara, CA, 2006, pp. 421–430.

[4] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, 2006.

[5] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[6] D. Lenat and R. Guha, *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Reading, MA: Addison-Wesley, 1990.

[7] H. Liu and P. Singh, "ConceptNet: A practical commonsense reasoning toolkit," *BT Technol. J.*, vol. 22, no. 4, pp. 211–226, 2004.

[8] A. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer, "Video content annotation using visual analysis and a large semantic knowledgebase," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, Madison, WI, 2003, vol. 2, pp. 327–334.

[9] A. Smeaton, "Large scale evaluations of multimedia information retrieval: The TRECVid experience," *CIVR*, vol. 3568, pp. 19–27, 2005.

[10] M. Brown, J. Foote, G. Jones, K. Sparck-Jones, and S. Young, "Automatic content-based retrieval of broadcast news," in *Proc. ACM Multimedia*, San Francisco, CA, 1995.

[11] B. Adams, A. Amir, C. Dorai, S. Ghosal, G. Iyengar, A. Jaimes, C. Lang, C.-Y. Lin, A. Natsev, M. Naphade, C. Neti, H. Nock, H. Permuter, R. Singh, J. Smith, S. Srinivasan, B. Tseng, T. Ashwin, and D. Zhang, "IBM research TREC-2002 video retrieval system," in *Proc. 11th Text Retrieval Conf.*, Gaithersburg, MD, 2002.

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

[13] W.-Y. Ma and B. Manjunath, "NeTra: A toolbox for navigating large image databases," *Multimedia Syst.*, vol. 7, no. 3, pp. 184–198, 1999.

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

[15] 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, 1998.

[16] T. Westerveld, A. de Vries, A. van Ballegooij, F. de Jong, and D. Hiemstra, "A probabilistic multimedia retrieval model and its evaluation," *J. Appl. Signal Process.*, vol. 2003, no. 2, pp. 186–197, 2003.

[17] T.-S. Chua, S.-Y. Neo, K.-Y. Li, G. Wang, R. Shi, M. Zhao, H. Xu, Q. Tian, S. Gao, and T. L. Nwe, "TRECVID 2004 search and feature extraction task by NUS PRIS," in *Proc. TRECVID Workshop*, Gaithersburg, MD, 2004.

[18] R. Yan, J. Yang, and A. Hauptmann, "Learning query-class dependent weights for automatic video retrieval," in *Proc. ACM Multimedia*, New York, 2004, pp. 548–555.

[19] A. Natsev, M. R. Naphade, and J. Tesic, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 598–607.

[20] L. S. Kennedy, A. Natsev, and S.-F. Chang, "Automatic discovery of query-class-dependent models for multimodal search," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 882–891.

[21] G. Iyengar, P. Duygulu, S. Feng, P. Ircing, S. Khudanpur, D. Klakow, M. Krause, R. Manmatha, H. Nock, D. Petkova, B. Pytlik, and P. Virga, "Joint visual-text modeling for automatic retrieval of multimedia documents," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 21–30.

[22] R. Lienhart, C. Kuhmünch, and W. Effelsberg, "On the detection and recognition of television commercials," in *IEEE Conf. Multimedia Computing and Systems*, Ottawa, ON, Canada, 1997, pp. 509–516.

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

[24] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. ACM Multimedia*, Los Angeles, CA, 2000, pp. 105–115.

[25] 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, 2001.

[26] 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.

[27] 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, 2004.

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

[29] J. C. van Gemert, J.-M. Geusebroek, C. Veenman, C. G. M. Snoek, and A. W. M. Smeulders, "Robust scene categorization by learning image statistics in context," in *Proc. Int. Workshop on Semantic Learning Applications in Multimedia*, New York, 2006.

[30] C.-Y. Lin, B. Tseng, and J. Smith, "Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets," in *Proc. TRECVID Workshop*, Gaithersburg, MD, 2003.

[31] M. Christel, T. Kanade, M. Mauldin, R. Reddy, M. Sirbu, S. Stevens, and H. Wactlar, "Informedia digital video library," *Commun. ACM*, vol. 38, no. 4, pp. 57–58, 1995.

[32] T. Volkmer, J. Smith, A. Natsev, M. Campbell, and M. Naphade, "A web-based system for collaborative annotation of large image and video collections," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 892–901.

[33] T. Volkmer, J. A. Thom, and S. M. M. Tahaghoghi, "Modelling human judgement of digital imagery for multimedia retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 967–974, Aug. 2007.

[34] L. Hollink, A. T. Schreiber, J. Wielemaker, and B. J. Wielinga, "Semantic annotation of image collections," in *Proc. K-Cap 2003 Workshop on Knowledge Markup and Semantic Annotation*, Sanibel Island, FL, 2003.

[35] L. Hollink, "Semantic Annotation for Retrieval of Visual Resources," Ph.D. Dissertation, Vrije Univ., Amsterdam, The Netherlands, 2006.

[36] E. Hyvönen, S. Saarela, K. Viljanen, E. Mäkelä, A. Valo, M. Salminen, S. Kettula, and M. Junnila, "A cultural community portal for publishing museum collections on the semantic web.," in *Proc. ECAI Workshop on Application of Semantic Web Technologies to Web Communities*, Valencia, Spain, 2004.

[37] A. T. Schreiber, B. Dubbeldam, J. Wielemaker, and B. J. Wielinga, "Ontology-based photo annotation," *IEEE Intell. Syst.*, vol. 16, no. 3, pp. 66–74, 2001.

[38] A. F. Smeaton and I. Quigley, "Experiments on using semantic distances between words in image caption retrieval," in *Proc. ACM SIGIR Conf. Research and Development in Information Retrieval*, Zurich, Switzerland, 1996, pp. 174–180.

[39] L. Hollink, M. Worring, and A. T. Schreiber, "Building a visual ontology for video retrieval," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 479–482.

[40] M. Bertini, A. Del Bimbo, and C. Torniai, "Automatic video annotation using ontologies extended visual information," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 395–398.

[41] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd 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] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua, "Video retrieval using high level features: Exploiting query matching and confidence-based weighting," in *CIVR, ser. LNCS,* H. Sundaram, *et al.,* Ed. Heidelberg, Germany: Springer-Verlag, 2006, vol. 4071, pp. 143–152.

[45] C. G. M. Snoek, M. Worring, D. C. Koelma, and A. W. M. Smeulders, "A learned lexicon-driven paradigm for interactive video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 280–292, 2007.

[46] G. Salton, "Dynamic document processing," *Commun. ACM*, vol. 15, no. 7, pp. 658–668, 1972.

[47] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[48] *"The Lucene Search Engine,"* 2006 [Online]. Available: http://lucene.apache.org/

[49] E. F. Tjong Kim Sang, "Memory-based shallow parsing," *J. Mach. Learn. Res.*, vol. 2, pp. 559–594, 2002.

[50] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Int. Joint Conf. Artificial Intelligence*, Montréal, QC, Canada, 1995, pp. 448–453.

[51] G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker, "A semantic concordance," in *Proc. Workshop on Human Language Technology*, Princeton, NJ, 1993, pp. 303–308.

[52] N. Rasiwasia, N. Vasconcelos, and P. J. Moreno, "Bridging the gap: Query by semantic example," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 923–938, Aug. 2007.

[53] M. F. Porter, "An algorithm for suffix stripping," in *Readings in Information Retrieval*. San Francisco, CA: Morgan Kaufmann, 1997, pp. 313–316.

[54] V. Hollink, J. Kamps, C. Monz, and M. de Rijke, "Monolingual document retrieval for European languages," *Inf. Retriev.*, vol. 7, no. 1-2, pp. 33–52, 2004.

[55] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness," *Comput. Linguist.*, vol. 32, no. 1, pp. 13–47, 2006.

[56] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 958–966, Aug. 2007.

[57] E. Fox and J. Shaw, "Combination of multiple searches," *TREC-2*, pp. 243–252, 1994.

[58] J. Kamps and M. de Rijke, "The effectiveness of combining information retrieval strategies for European languages," in *Proc. 19th Annu. ACM Symp. Applied Computing*, 2004, pp. 1073–1077.

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 with the Intelligent Systems Laboratory, University of Amsterdam. He has published in more than 40 scientific publications.

Dr. Snoek is the local chair for ACM CIVR 2007 and Lead Researcher of the MediaMill semantic video search engine.

Bouke Huurnink received the M.Sc. degree in information science, majoring in multimedia systems, from the University of Amsterdam, Amsterdam, The Netherlands, in 2005. He is currently pursuing the Ph.D. degree at the Intelligent Systems Laboratory, University of Amsterdam, with a research focus on multimedia information retrieval.

Laura Hollink received the M.A. degree in social science informatics from the University of Amsterdam, Amsterdam, The Netherlands, in 2001. In 2006 she received the Ph.D. degree in computer science from Vrije University, Amsterdam, on the topic of semantic annotation for retrieval of visual resources.

Currently, she is a Postdoctoral Researcher at the Intelligent Information Systems Group, Vrije University. She is Daily Project Manager of the NWO MuNCH Project and works part-time at the Dutch Institute for Sound and Vision in Hilversum.

Maarten de Rijke is Professor of information processing and Internet at the University of Amsterdam, Amsterdam, The Netherlands. He has published over 350 papers and books. He is the local organizer for SIGIR 2007 and founder of ZookMa, a startup aimed at providing insight in Dutch language consumer generated content.

Guus Schreiber is a Professor of intelligent information systems at the Free University, Amsterdam, The Netherlands. He has published 140 articles and books. In 2000, he published with MIT Press a textbook on knowledge engineering and knowledge management, based on the CommonKADS methodology.

Dr. Schreiber is co-chairing the W3C Semantic Web Deployment Working Group and is the former co-chair of the Web Ontology and the Semantic Web Best Practices Groups. He is also Scientific Director of the IST Network of Excellence "Knowledge Web."

Marcel Worring (M'03) is an Associate Professor in computer science at the University of Amsterdam, Amsterdam, The Netherlands. He is leading the MediaMill team, which has participated in all TRECVID editions and has published over 100 scientific publications in journals and proceedings.

He is an associate editor for the IEEE TRANSACTIONS ON MULTIMEDIA and is the chair of the IAPR TC12 on Multimedia and Visual Collections.