Assessing Concept Selection for Video Retrieval

Bouke Huurnink
bhuurnin@science.uva.nl

Katja Hofmann
khofmann@science.uva.nl

Maarten de Rijke
mdr@science.uva.nl

ISLA, University of Amsterdam
Kruislaan 403, Amsterdam
The Netherlands

ABSTRACT

We explore the use of benchmarks to address the problem of assessing concept selection in video retrieval systems. Two benchmarks are presented, one created by human association of queries to concepts, the other generated from an extensively tagged collection. They are compared in terms of reliability, captured semantics, and retrieval performance. Recommendations are given for using the benchmarks to assess concept selection algorithms; the assessment is demonstrated on two existing algorithms. The benchmarks are released to the research community.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [**Information Systems Applications**]: H.4.2 Types of Systems; H.4.m Miscellaneous

General Terms

Measurement, Performance, Experimentation

Keywords

Video retrieval, Concept selection

1. INTRODUCTION

Video collections are becoming widely available, raising the need for effective access to video content. One way to facilitate access is concept-based video retrieval, where visual concepts are detected in video. A recent trend in concept-based video retrieval has been to search for generic methods that learn to detect concepts using examples [8, 18, 25]. Given a video collection with concept annotations, we can use a retrieval approach similar to text retrieval, where concepts are considered textual labels that can be indexed and retrieved using a standard text retrieval engine (cf. Figure 1). A difference with text retrieval is that users' textual queries have to be translated into visual concepts, a step called concept selection.

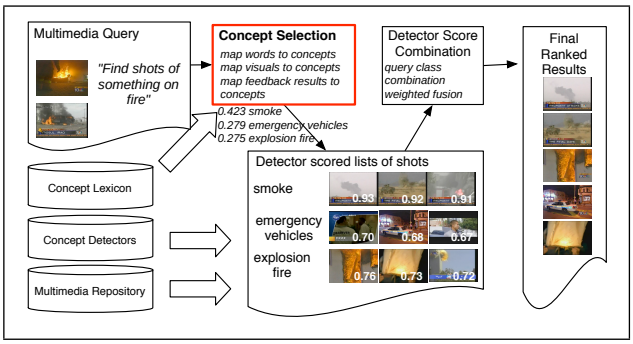

Figure 1: Outline of a concept detector-based retrieval system, adapted from [9]. For simplification, other possible sources of information for retrieval, such as speech and low-level features, are omitted.

Developing concept selection algorithms has so far been difficult. Often, these algorithms are designed to match semantically relevant concepts to a query, but can only be assessed *extrinsically*, in the context of a full video retrieval system (see Section 2 for references). The uncertain detection of concepts makes it difficult to assess the semantic relevance of a concept to a query.

We address the problem of evaluating concept selection algorithms, explore methods of creating re-usable benchmarks, and determine ways of using them to measure the quality of a (system-based) concept selection method. The use of these benchmarks allows us to evaluate concept selection independently from other components of video retrieval systems. We propose two benchmarks for assessing concept selection, one human-generated, the other back-generated from a video collection annotated with concepts. For the human benchmark, focus group experiments are used to identify visual concepts that the participants would consider useful to answer a query. For the collection benchmark, an extensively labelled video collection is used to determine which concepts are best for a query. After introducing and analyzing the benchmarks we demonstrate how they can be used to assess concept selection algorithms. To illustrate the use of our benchmarks and metrics, we assess two concept selection methods proposed in the literature. We determine

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.

how the selected concepts relate to end-to-end retrieval performance in the context of an 'oracle' multimedia retrieval system, where visual concepts are perfectly detected. We develop recommendations as to when and how the benchmarks can be used to assess concept selection algorithms. The human and collection concept selection benchmarks that we describe are being made available to the research community, together with a script that implements our evaluation measures.[1]

The rest of this paper is structured as follows. Related work is discussed in Section 2. Section 3 presents our methodology and development of two benchmarks, which are compared in Section 4. We investigate the application of the benchmarks to the assessment of concept selection in Section 5. Discussion and practical recommendations are given in Section 6, and Section 7 presents conclusions and identifies areas for future work.

2. RELATED WORK

Several areas relate to the evaluation of concept selection for video retrieval. We first outline the collaborative video annotation efforts typically used to evaluate video retrieval systems. Next, we give an overview of automatic concept selection methods and previously used approaches to assessing concept selection for video retrieval.

2.1 Collaborative video annotation

Video retrieval is typically evaluated on shared benchmark collections and shared annotation efforts. Efforts by TRECVID [17], LSCOM [8], and MediaMill [18] to create shared pools of training data have enabled researchers to build video retrieval systems equipped with detectors for large numbers of semantic concepts. In addition, TRECVID and LSCOM have both released truth annotations for a number of video retrieval queries, in the form of topics or use case queries (for ease of use we will refer to both as topics). The combination of topics and ground truth annotations are widely used to compare different systems, and assess new retrieval algorithms [16].

2.2 Concept selection methods

We identify three broad approaches to concept selection: automatic, human association, and generation from extensively annotated video data.

2.2.1 Automatic concept selection

We use the term 'automatic concept selection' to describe the concept selection algorithms used in video retrieval systems to automatically translate a query to the system concept lexicon, usually returning a weighted list of concepts as a result. There is much work being done in this area; we give a short review to indicate the scope of the research. Natsev et al. [9] divide concept selection algorithms, which they term 'concept-based query expansion approaches,' into three broad categories: text-based, visual-based, and result-based. Many detector-enabled video retrieval systems use a number of such techniques, see e.g., [9, 10, 19, 24].

Text-based algorithms match the query text to the concepts in the lexicon, e.g., against textual descriptions of the concepts, or through computing lexical similarity using resources such as WordNet [7]. Visual-based algorithms use similarity of detector models to visual examples to identify appropriate concepts. Result-based algorithms use feedback from an initial retrieval step to select and weight concepts.

2.2.2 Utilising human associations

Knowledge about the world is implicit in human minds, and retrieval systems can exploit this knowledge by asking humans to select appropriate concepts for individual queries. Christel and Hauptmann [1] analysed such associations, with concept usefulness to a query rated on a scale of 1–5. Two collections were used, one containing 23 queries and 10 concepts, the other containing 24 queries and 17 topics. Their work showed inter-annotator agreement to be low, with less than 15% of the mappings being agreed upon by all participants. Neo et al. [10] found agreement to be similarly low in an experiment with 8 queries and 24 concepts. To compare, inter-annotator agreement for users assigning concept labels to video fragments can be higher than 95% [22].

2.2.3 Utilising labelled video collections

Recent efforts to label large video collections have allowed researchers to investigate generative concept selection, using data sets manually annotated with respect to both queries and concepts [1, 3]. Relevant concepts are selected by analysing which shots are relevant to a query, and in turn which concepts are associated with relevant shots. This can be done by, e.g., using mutual information to determine the utility of a concept [6]: a concept is mapped to a query if it reduces uncertainty about a particular shot being relevant to that query. Another approach is to compute the probability a shot is relevant to a query given that the shot is associated with a particular concept, normalising for prior probability that any shot is relevant to the query [1].

2.3 Assessing concept selection in a video retrieval setting

Content-based video retrieval systems use a range of techniques to select concepts that are relevant for a set of trained detectors. The most successful systems are those in which an expert user manually selects the most appropriate concepts for a query [17]. Such systems demand a high user investment in the retrieval task, and there is recognition that video retrieval systems should be easier to use to be successful [16]. Therefore *automatic* retrieval systems that do not require user interaction have received much attention. In such systems, appropriate concepts be automatically derived from the original (multimedia) query. Automatic selection algorithms are often assessed within the context of an end-to-end multimodal retrieval system [9, 10, 20, 24]. This assessment is often not explicit, but judges automatic selection together with other system components such as text search, search-by-example, and relevance feedback methods.

Others have isolated the detectors, creating detector-based retrieval systems that allow for closer examination of the impact of concept selection on retrieval performance. E.g., [19, 23, 24] use text- and image-based algorithms to select the one best concept for a query. Evaluation is then done by ranking video fragments according to the scores assigned by the associated concept detector, and assessing against a pooled truth. Another approach is to combine detector scores from multiple selected concepts [23, 24].

[1]See `http://ilps.science.uva.nl/Resources/ConceptSelectionBenchmarks/`.

Finally, human judgements can be used to assess concept selection independently of detector performance. Neo et al. [10] assess system concept selection (for 8 queries) by comparing the agreement with a set of human association judgments. They find low agreement but do not investigate correlations with final retrieval performance.

3. BENCHMARK DEVELOPMENT

In the previous section we identified two non-automatic approaches for concept selection, one utilizing *labeled video* and the other utilizing *human associations*. We develop benchmarks corresponding to each approach. First, the *collection benchmark* is back-generated from labeled video. Second, the *human benchmark* is based on a user experiment where human subjects associate concepts with given topics.

3.1 Collection benchmark

Labeling video collections with truth judgments for hundreds of concepts and tens of topics requires large scale annotation efforts. Once annotations are completed, they can be used to deduce which concepts are relevant to a topic.

3.1.1 Methodology

To specify our collection benchmark we follow Hauptmann et al. [3] and employ the information-theoretic notion of mutual information. Denote relevance of a shot for a given topic as T and the presence or absence of a concept in a shot as C. Both T and C are binary random variables. The mutual information between T and C is then defined as:

$$I(T;C) = \sum_{t,c} P(t,c) \log \frac{P(t,c)}{P(t)P(c)} \quad (1)$$

with $t \in \{\texttt{relevance}, \texttt{irrelevance}\}$, $c \in \{\texttt{presence}, \texttt{absence}\}$; estimates are derived from a tagged collection [6]. Concepts are ranked according to how much $I(T:C)$ reduces the entropy of T using maximum likelihood estimates.

To define the concept selection that constitutes the collection benchmark, we impose two restrictions on this ranked list. First, we use the suggested threshold of 1% to eliminate concepts with low mutual information. Second, we remove negatively correlated concepts as follows: recall that mutual information assigns high scores to both positively and negatively correlated variables [2]. Negatively correlated concepts are identified using *pointwise mutual information*:

$$I(t;c) = \log \frac{P(t,c)}{P(t)P(c)}. \quad (2)$$

If $I(\texttt{absence};\texttt{relevance})$ of a concept is greater than $I(\texttt{presence};\texttt{relevance})$ we discard it from the mapping.

3.1.2 Development data

To create the collection benchmark we require topics for which appropriate concepts are to be selected, a lexicon from which to select concepts, and a video collection in which shots are annotated with relevant concepts and topics. The topics and concept lexicon we can use depend heavily on the annotations available in the video collection.

For our video collection we use the development data set from the TRECVID 2005 corpus [11], which consists of about 70 hours of English, Chinese, and Arabic news video from October and November of 2004, and has been automatically segmented into over 40,000 shots. This is the only collection that we use, and we split it into a training and a test set. Following the TRECVID strategy [11], we split the video collection in half chronologically by source.

The MediaMill and LSCOM efforts have released lexicons of 101 and over 400 concepts, respectively, and produced extensively annotated truth data for the TRECVID 2005 development data that we use. We combine the two lexicons using disambiguation data from [19]. The final combined lexicon consists of 450 concepts. The video collection has also been annotated with respect to 50 LSCOM use case queries[2] and the 24 TRECVID 2005 test topics,[3] resulting in a combined set of 74 topics. Only 52 of these have relevant shots in the training collection; these are the topics included in our collection benchmark.

3.2 Human benchmark

People can have a wide range of associations with a concept, depending on context and personal characteristics. Nevertheless, there exists a common understanding of concepts that is socially constructed and allows people to communicate [5, 15]. The goal of the human-generated benchmark is to capture this common understanding, as opposed to the wider range of individual associations [3, 10].

3.2.1 Methodology

We conducted two focus group experiments following the same procedure. The first was conducted in January 2008 and the second in March 2008, with respectively 3 and 4 undergraduate students. Both studies consisted of several sessions held over a two week period. None of the subjects had prior experience retrieving video with detectors.

Each study consisted of two phases. The first was designed to familiarize subjects with the concept lexicon. 450 flash cards, each printed with the name and description of a concept, were given to the group. The subjects were asked to collaboratively organize the cards so that relevant concepts for a topic could be identified quickly. The second phase was aimed at determining the relevant concepts for different topics. Subjects were given 8 topics at a time and asked to note down concepts that would help them find relevant shots in a video collection. Next, the concepts noted by each subject were collected on a flip board, and subjects were asked to vote on which concepts they thought would help them find relevant shots. During this group process discussion was encouraged. As a result, concepts in the list might receive no votes. Finally, subjects were asked to unanimously decide which concept would be best for retrieving relevant shots.

To allow comparison between the two studies, they included 30 overlapping topics. These were randomly selected from the overall list of topics outlined in the next section, and included in the same positions in each study, i.e. if the third topic of group 1 was an overlapping topic, it was also positioned as the third topic for group 2. We did this because we assume that order and learning effects play a role; studying these effects is beyond the scope of this paper.

3.2.2 Development data

To create the human benchmark we require topics for which to select appropriate concepts, and a concept lexicon from which the concepts may be drawn. All topics from the

[2] `http://www.lscom.org/useCaseQueries/index.html`
[3] `http://www-nlpir.nist.gov/projects/tv2005/topics/`
Annotations were kindly donated to us by CMU [26].

TRECVID 2005 and LSCOM sets described in Section 3.1 are used to create the human benchmark. This allows for comparison between the human benchmark and the collection benchmark. In addition, the topics from the TRECVID 2006 and 2007 benchmarks are included, as these are widely used for evaluating video retrieval systems [16]. This results in a total of 122 topics. We utilize the concept lexicon of 450 concepts described in Section 3.1, again to facilitate comparison. All concepts in the lexicon are accompanied by descriptions created to aid annotators.

4. BENCHMARK EVALUATION

In this section we compare the two concept selection benchmarks along three dimensions:

Reliability: the method used for creating the benchmark data must be reasonably robust.

Semantics: concept selection should capture meaningful relationships between topics and concepts.

Retrieval performance: the benchmark data should perform well for concept-based video retrieval.

We expect each benchmark to capture different aspects of the concept selection task and as a result show different behavior regarding our evaluation criteria.

4.1 Reliability

4.1.1 Human agreement

Reliability of the human benchmark is assessed by analyzing agreement between user studies 1 and 2 for the 30 topics annotated by both groups. When choosing the best topic, the groups agreed on 80% of the topics. In cases where the best topic did not match, the best topic of one group was still ranked high by the other group.

Because of the difference in the number of concepts, we compare the two sets of selected concepts using asymmetric set difference. In one case this results in 77.5% agreement, i.e., the group in study 1 identified 77.5% of the concepts that were selected by the group in study 2. In the other case there is an overlap of 34%—group 2 found 34% of the concepts found by group 1.

On average, group 1 considered 7.5 concepts per topic relevant, while group 2 selected 4.6 concepts per topic. Group 2 both identified fewer concepts during the individual selection round, and removed more concepts during the voting round. This difference is attributed to: (1) group size—the group in study 1 had four members, while the group in study 2 only had three, and (2) the physical layout of the concept hierarchy—in study 1, subjects ensured all concepts were visible during the selection process, while in study 2 subjects created a layout where similar concepts overlapped.

4.1.2 Varying labelled collection size

The collection benchmark may be reproduced by anyone with the annotated video material. But is it reliable across collection sizes? What if a smaller collection is annotated? To assess this, we generated the benchmark using the first 10% of the training data, selected chronologically. We compare the resulting data to the data created when using the first 20%, 30%, ..., 100% of the training data and evaluate this against the human benchmark using asymmetric set overlap, i.e., the number of human selected concepts identified by the collection benchmark. We restrict our analysis to the 23 topics with relevant shots in the first 10% of the training set.

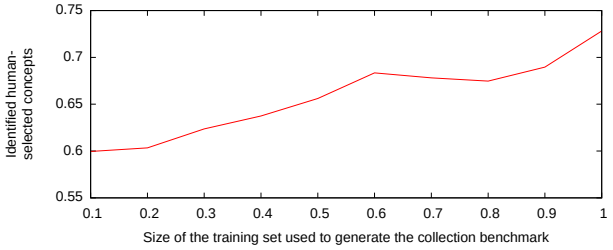

Figure 2: Proportion of human-assigned concepts identified by the collection benchmark, using different training set sizes.

Figure 2 shows the concept overlap as the size of the generation data is altered. The proportion of human-selected concepts slowly increases from 60% to 73%, depending on the amount of generation data. Even with a small amount of data the collection benchmark captures a substantial portion of the relationships between topics and concepts considered meaningful by the human subjects.

4.1.3 Collection vs. human

Overall, the collection benchmark selected more concepts per topic and selected a wider range of topics than the human benchmark. On average, 25.3 concepts were selected

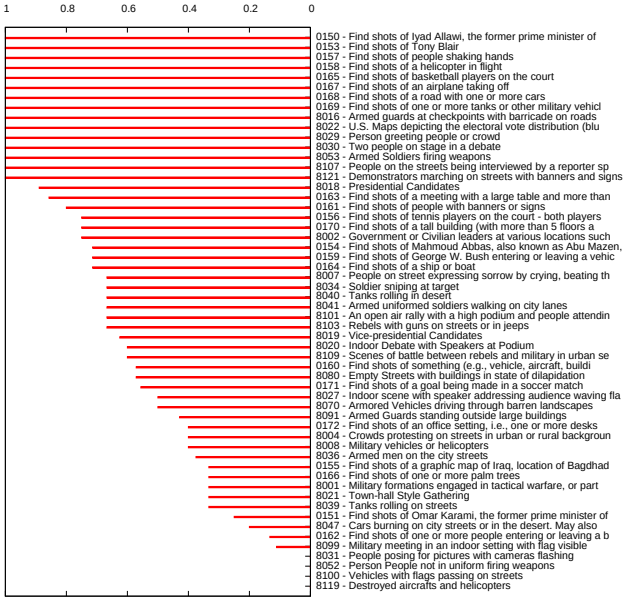

Figure 3: Per-topic proportion of human-assigned concepts identified by the collection benchmark.

per query, compared to 5.8 concepts selected by the human subjects. Out of the available 450 concepts, the collection benchmark contains 268 unique concepts , while the human benchmark contains 189 unique concepts.

The average proportion of human-selected concepts also selected by the collection benchmark is 63%. Figure 3 shows

Table 1: Most frequently occurring concepts.

Human benchmark		Collection benchmark	
count	concept description	count	concept description
13	Armed Person	36	Daytime Outdoor
11	Riot	35	Outdoor
9	Violence	32	Person
8	Street Battle	31	Adult
8	Military Ground Vehicle	24	Sky
8	Insurgents	24	People Walking
8	Fire Weapon	22	Violence
7	Non-uniformed Fighters	21	Weapons
7	Military Personnel	20	Civilian Person
7	Head Of State	19	Vehicle
7	Demonstration Or Protest	19	Road
7	Crowd	18	Military Personnel
6	Government Leader	18	Face
6	Election Campaign Greeting	17	Politics
5	Urban	17	Machine Guns
5	Speaker At Podium	16	Soldiers
5	Soldiers	16	Armed Person

the overlap per topic. For many topics, the collection benchmark contains many of the human-selected concepts. For some topics however, the agreement is low, and for four topics no concepts overlap.

4.2 Semantics

To compare the semantics captured by the two benchmarks we compare frequently selected concepts, and analyze example topics where we observe large differences between selected concepts. Table 1 shows the most frequently selected concepts for both human and collection benchmark. The collection benchmark frequently selects general concepts such as *Daytime Outdoor*, *Outdoor*, or *Person*. In contrast, the human benchmark selected *Daytime Outdoor* only once and never selected *Outdoor* or *Person*. Concepts frequently selected by the human benchmark, such as *Armed Person*, *Riot*, *Violence*, and *Street Battle* are frequently selected by the collection benchmark as well.

Human subjects tend to select a small number of highly relevant concepts, while the collection benchmark also selects broader and more loosely associated concepts (see Table 2). For example, the topics *0150 – Find shots of Iyad Allawi*, and *0153 – Find shots of Tony Blair*, both human and collection benchmarks selected highly precise concepts, such as *Iyad Allawi*, *Tony Blair*, and *Head of State*. The collection benchmark added broader concepts, such as *Male Person* or *Politics*. Concepts like these were considered to be 'too broad' by the study subjects creating the human benchmark. While these concepts were considered relevant to the topic, subjects mentioned that these concepts applied to too many shots and would therefore not be useful when retrieving shots for a narrower topic.

The collection benchmark also selects concepts that are not directly related but co-occur with a topic in the specific collection. For example, topic *0166 – Find shots of one or more palm trees* is associated with *Weapons*, *Fire Weapon*, and *Shooting*. This reflects the fact that we are working with a collection of news broadcasts, with many armed conflicts taking place in tropical countries. Other co-occurrences may be random artifacts that do not necessarily reflect collection characteristics. For example, for topic *0150 – Find shots of Iyad Allawi* the collection benchmark selected the concept *Outer Space*, which proved to be the result of a small number of shots with problematic shot boundary detection.

For a small number of concepts the collection benchmark did not select any of the human-selected concepts. For topic *8100 – Vehicles with flags passing on streets*, the human benchmark appears to focus mostly on the fact that flags should be visible. The collection benchmark selected concepts related to vehicles, but none related to flags. For topic *8119 – Destroyed aircrafts and helicopters*, the human benchmark selected concepts related to aircraft and helicopters while the collection benchmark did not. In these cases two factors may play a role. First, the number of relevant shots for these topics is relatively small (respectively, 8 and 9 shots). Although we did not find a correlation between benchmark agreement and the number of relevant shots, a minimal number of representative shots may be required to select meaningful concepts for the collection benchmark. Second, the associated concepts may be difficult to judge in a shot relevant to these topics. When small flags are shown on a car, a human judge may not mark the concept *Flag* in this shot, and shots of destroyed aircraft may not be annotated with the concept *Aircraft*.

4.3 Retrieval performance

We assess retrieval performance with an 'oracle' multimedia retrieval system, in which concepts are perfectly detected. This allows us to bypass the problem of uncertainty faced by video retrieval systems that rely on trained detectors to identify concepts in video shots.

In keeping with our text retrieval approach, we model concepts as terms. Each shot is represented by a list of concept terms. We index the shots from the collection described in Section 3.1, and use the well-established BM25 [13] formula to retrieve shots that are most relevant to the query concepts. We do not incorporate the weightings available in our benchmark into the retrieval queries, as we wish to make as few assumptions as possible about our system. Retrieval results are evaluated against truth judgments in terms of Average Precision (AP). We limit evaluation to the 52 topics for which relevant shots have been identified in both the training (used to generate the collection benchmark) and test collection (used to evaluate retrieval of both benchmarks).

We expect the collection benchmark to outperform the human benchmark, as it is tuned to the collection domain of retrieval broadcast news. This is confirmed by the retrieval results; the Mean Average Precision (MAP) score for retrieval using the human benchmark is 0.056, while for retrieval with the collection benchmark it is 0.204.

Looking at the per-topic differences in AP we note that the human benchmark outperforms the collection benchmark in a small number of cases (Figure 4). For topic *0153 – Find shots of Tony Blair* the perfectly overlapping concept *Tony Blair* identifies all shots relevant to this topic. The collection benchmark introduces less related concepts that reduce precision and hurt retrieval performance. Similarly, the concepts selected in the human benchmark for topic *0156 – Find shots of tennis players on the court - both players visible at same time* appear to be precise and able to retrieve most relevant shots. Additional concepts such as *Athletes*, or *Sports Venue* improve recall but hurt precision.

The concepts associated by human subjects typically favor precision, while the collection generated ones favor recall. This is a result of how the annotated collection is produced. Relevance judgments for video retrieval appear to be recall oriented, i.e., a shot is judged relevant even if the topic only appears in a small part of the shot. The human subjects appear to aim for high precision.

In sum, the collection benchmark captures meaningful relationships between topics and concepts, but also includes artifacts. Comparing benchmark data generated in such a way to human judgments can give insight into collection

Table 2: Example topics and concepts associated by human and collection benchmark.

topic	topic description	human benchmark	collection benchmark
0150	Find shots of Iyad Allawi, the former prime minister of Iraq	Iyad Allawi, Head Of State	Iyad Allawi, Non-us National Flags, Government Leader, Press Conference, Interview On Location, Male News Subject, Address Or Speech, Speaker At Podium, Microphones, Ties, Politics, Pope, Adult, Meeting, Person, Face, Head And Shoulder, Single Person Male, Head Of State, Conference Room, Male Person, Suits, Individual, Outer Space, Soldiers, Civilian Person, Talking
0153	Find shots of Tony Blair	Politics, Male News Subject, Tony Blair, Head Of State	Tony Blair, Politics, Head Of State, Ties, Speaker At Podium, Address Or Speech, Standing, Single Person Male, Face, Civilian Person, Flag USA, Flag, Adult, Single Person, Male Person, Non-us National Flags, Person, Individual, Government Leader, George Bush jr, Speaking To Camera, Furniture, Male News Subject
0156	Find shots of tennis players on the court - both players visible at same time	Daytime Outdoor, Running, Sport Games, Tennis Game	Tennis Game, Sport Games, Athlete, Grandstands Bleachers, People Walking, Running, Nighttime, Indoor Sports Venue, Entertainment, Walking Running, Stadium, Caucasians, Person, Celebrity Entertainment, Adult, Overlayed Text, Scene Text
0166	Find shots of one or more palm trees	Forest, Vegetation, Desert, Tropical Settings, Tree, Beach	Weapons, Fire Weapon, Shooting, Outdoor, Daytime Outdoor, Sky, Military Personnel, Rifles, Armed Person, Street Battle, Soldiers, Machine Guns, Violence, Ground Combat, People Walking, Vegetation, Tree, Backpack, Backpackers, Mountain, Military Base, Ground Vehicles, Building, Rocky Ground, Ruins, Sunny, Tanks, Adobehouses, Canal, Smoke, Cul de sac, Cart Path, Mosques, Power Transmission Line Tower, Urban
8100	Vehicles with flags passing on streets	Non-us National Flags, Flag, Flag USA, Military Ground Vehicle	Pickup Truck, Police Security Personnel, Truck, Ground Vehicles, Car, Vehicle, Road, Sky, People Walking, Adobehouses, Dirt Gravel Road, Scene Text, Demonstration Or Protest, Armed Person, Standing, Outdoor, Daytime Outdoor, Vegetation, Building, Face, Windows, Person
8119	Destroyed aircrafts and helicopters	Fighter Combat, Aircraft, Helicopters, Emergency Vehicles, Airplane Crash	Ruins, Demonstration Or Protest, Entertainment, Daytime Outdoor, Overlayed Text, Windows

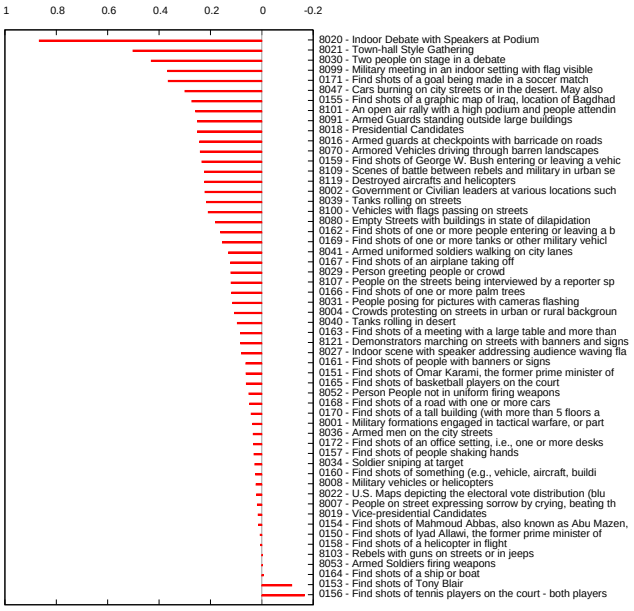

Figure 4: Per-topic difference in AP scores between human and collection benchmark.

characteristics that may be important for video retrieval. The collection benchmark is most successful at retrieval in the tested instance of a broadcast news collection; however, the human benchmark captures more general concept associations and may be more useful to test selection algorithms designed to be generalizable across video collections.

5. USING THE BENCHMARKS TO ASSESS CONCEPT SELECTION

In this section we examine scoring methods for predicting whether a set of automatically selected concepts will perform well for retrieval. We first describe two benchmark scoring measures, one based on set agreement and the other on rank agreement with the benchmark. We then examine the predictive power of the scoring methods when assessing concept selection for video retrieval, with perfect detectors.

5.1 Benchmark scoring measures

Consider a set of benchmark concepts C_B and a set of candidate concepts C_S, selected from a concept lexicon $\mathcal{C}$ so that $C_B \subset \mathcal{C}$ and $C_S \subset \mathcal{C}$. We assign two benchmark scores $score(C_B, C_S)$, using an increasing amount of information:

set agreement $score_{sa}(C_B, C_S)$ is defined as set agreement, the positive proportion of specific agreement between C_B and C_S [4]: $score_{sa}(C_B, C_S)$ is equal to 1 when $C_B = C_S$, and 0 when $C_B \cap C_S = \emptyset$. Ranking information is not included in this measure.

rank correlation $score_{rc}(C_B, C_S)$ is given by Spearman's rank correlation [21]. The $score_{sa}(C_B, C_S)$ is equal to 1 when $C_B = C_S$, and -1 when $C_B \cap C_S = \emptyset$. This measure takes ranking agreement into account.

5.2 Test case: Assessing concept selection

To assess the scoring methods and benchmark for the task of concept selection, we implement two automatic concept selection algorithms from the literature. Our aim is to review the ability of a benchmark to predict concept selection performance, rather than extensively reviewing concept selection approaches. We perform automatic concept selection for all 52 test queries. We assess the scores of the concept selection algorithms against both benchmarks, and analyze the predictive power of the benchmarks for concept-based video retrieval. We do this by contrasting the benchmark scores with the retrieval performance of the concept selection algorithms, using the retrieval setup of the 'oracle' retrieval system with perfect detectors outlined in Section 4.3.

5.2.1 *Automatic concept selection algorithms*

We implement two concept selection algorithms, based on *text matching* and on *ontology querying*. The text matching algorithm matches query text to a concept's textual description (after stop word removal), as done in [19, 23]. Concepts are ranked according to their similarity to the original query using the vector space model [14]. All returned concepts are added to the concept selection set. The ontology querying algorithm assigns query terms to nouns in WordNet [7], following [19]. The query nouns are related to concepts, which are also assigned to WordNet nouns. Concepts are assigned scores using Resnik's measure of information content [12]. To prevent this algorithm from returning all concepts in the lexicon we only select concepts with an information content higher than 5.

5.2.2 *Benchmark scores*

A preliminary investigation of the scores produced by the benchmarks is shown in Figure 5. These are new benchmarks, so we provide a small exploratory analysis. Note that we do not use averaged values, as they are not necessarily meaningful for measures such as set agreement and

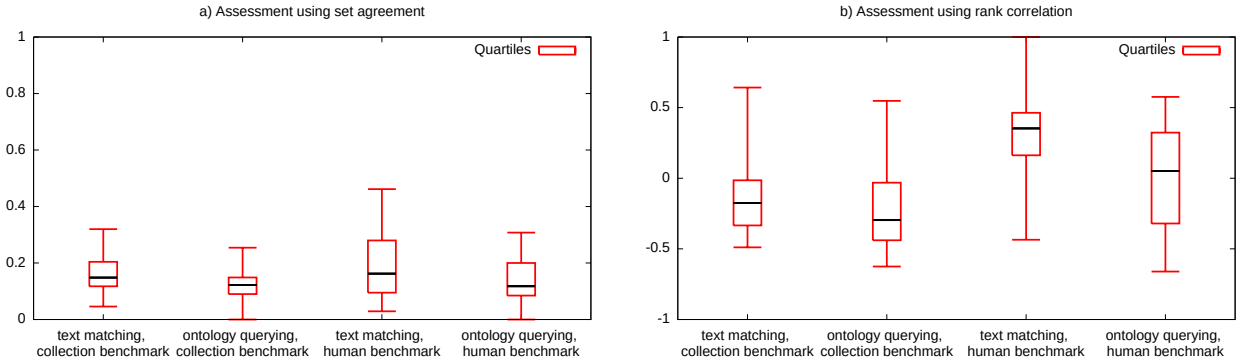

Figure 5: Assessing selection approaches. Text matching achieves higher scores on both benchmarks.

ranked correlation. Text matching results in higher median and quartile scores than ontology querying, no matter which scoring measure or benchmark is used. The benchmarks tend to agree overall on which concept selection strategy should be assigned the highest scores. This need not indicate agreement on individual topics, as we discuss below.

5.2.3 *Predicting video retrieval performance?*

The benchmark scores indicate that the concepts selected by text matching are better than those selected by ontology querying. Logically, then, text matching should produce better video retrieval results. This is confirmed by the final evaluation scores, with text matching giving an overall MAP of 0.043 and ontology querying giving a MAP of 0.015.

Table 3: Overall accuracy of the benchmarks in predicting the best retrieval performance.

	Retrieval prediction accuracy	
Score type	**Collection**	**Human**
Set agreement	67%	67%
Rank correlation	44%	73%

We also investigate prediction accuracy at the topic level (Table 3).[4] The different combinations of scoring measure and benchmark vary in accuracy when predicting which concept selection algorithm will give the best retrieval results. The best topic-level predictions of the human benchmark are achieved using the set overlap measure, while rank correlation in combination with the collection benchmark gives the best overall prediction results. This difference may be due to the different kinds of ranking information contained in each benchmark; only limited ranking information is available for the human benchmark, as ranking is based on the number of votes assigned by focus groups participants. This results in many concepts having the same rank. In contrast, the collection benchmark contains fine-grained rankings due to the underlying mutual information scores.

6. DISCUSSION

Having illustrated the development, evaluation, and application of two benchmarks for assessing concept selection, we summarize our observations and give recommendations for using the benchmarks to assess concept selection methods.

[4]Per-topic predictions are computed by determining which concept selection algorithm scores best against the benchmark. If the same selection algorithm performs best in terms of AP, the prediction is considered to be accurate.

6.1 Assessing our benchmarks

Table 4 summarizes our assessment of the human and collection benchmarks according to Section 4's criteria.

Table 4: Assessing the benchmarks.

Human benchmark	**Collection benchmark**
Reliability	
Repeated group experiment achieved 78% overlap and agreement on best concept on 80% of the topics.	Can be perfectly reproduced with same collection, resilient to changes in amount of training data.
Semantics	
Captures human world knowledge, shared understanding of topic and concept within a group of people.	Captures some world knowledge, but also collection-specific associations and noise.
Retrieval performance	
Reasonable performance on ideal collection, can predict performance of a concept selection method in retrieval task with reasonable accuracy.	Tuned to collection, excellent performance on ideal collection, can predict concept performance of a concept selection method in retrieval task with high accuracy.

Two additional observations may be of interest to potential users of our benchmark creation methodology. First, when *adding new topics* one can use the same setup and same concept lexicon when using the human benchmark methodology; for the collection benchmark relevance judgments over the entire collection are required for the new topics. Second, when *adding new concepts* to the human benchmark the focus group effort must be repeated with the new set of concepts. For the collection benchmark, one must annotate the video collection with respect to the new concepts (while retaining existing annotations).

6.2 Using the benchmarks to evaluate concept selection algorithms

Next, we give recommendations to assist users in choosing a benchmark for the assessment of a new concept selection algorithm. This choice depends on the goals one has. If one wishes to capture general world knowledge, and form collection-independent associations, the human generated benchmark is likely to be more appropriate. In contrast, if one wishes to capture more collection-specific associations (specifically in the broadcast news domain), the collection benchmark is is more appropriate. The human benchmark rewards precise selection methods that return few concepts,

while the collection benchmark is much more recall-oriented, rewarding methods that return many concepts.

Our final recommendation concerns the choice of metric to be used for assessment. We recommend that set overlap be used when scoring concept selection against the human benchmark, and rank correlation when scoring concept selection against the collection benchmark (see Section 5).

7. CONCLUSIONS AND FUTURE WORK

We have isolated the task of assessing automatic concept selection algorithms for video retrieval. This paper is a first step in assessing concept selection independently of detector performance. By using knowledge external to the system, we have developed two benchmarks for scoring selection algorithms, each benchmark capturing a different type of knowledge. Both benchmarks consist of a set of queries. Every query is mapped to concepts from a lexicon of 450 visual concepts according to either collection knowledge or human association. We examined the methodology used to create each benchmark, and considered the implications for assessing concept selection methods using the benchmarks. As developing benchmarks requires a considerable investment, we release our benchmarks to the research community.

We demonstrated the use of our benchmarks for automatic concept selection by introducing two ways of scoring a set of concepts: one based on set agreement with the benchmark, the other incorporating rank correlation. In a test case, we scored two automatic concept selection methods against the benchmarks. We assessed their retrieval performance against an oracle video retrieval system, with 'error-free' concept detection and simple text-based retrieval using the selected concepts as a query. Here we found that both benchmarks could be used to predict which concept selection algorithm would give the overall best performance.

The release of the benchmarks opens up several directions for future research. Many concept selection algorithms have been developed, and the benchmarks provide a platform for reviewing whether they select concepts in a semantically meaningful way. The benchmarks could also be used to develop concept suggestion algorithms that aid users of interactive video retrieval systems. Finally, we plan to investigate the problem of detector combination, using the benchmarks to determine the detectors to be combined.

8. ACKNOWLEDGMENTS

Huurnink and De Rijke were supported by the Netherlands Organization for Scientific Research (NWO) under project number 640.002.501. Hofmann and De Rijke were supported by NWO under number STE-05-039. De Rijke was also supported by NWO under numbers 017.001.190, 220-80-001, 612.066.512, 640.001.501, STE-07-012, and by the E.U. IST program of the 6th FP for RTD under project MultiMATCH contract IST-033104. We thank Cees Snoek for useful discussions and comments, and our students for their participation in the user studies.

References

[1] M. G. Christel and A. G. Hauptmann. The use and utility of high-level semantic features in video retrieval. In *CIVR '05*, pages 134–144, 2005.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[3] A. G. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. D. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Trans. Multimedia*, 9(5):958–966, 2007.

[4] B. L. Joseph L. Fleiss and M. C. Paik. The measurement of interrater agreement. In *Statistical methods for rates and proportions*, pages 598–626. John Wiley & Sons, 2004.

[5] G. Lakoff. *Women, Fire, and Dangerous Things*. University Of Chicago Press, 1990.

[6] W.-H. Lin and A. G. Hauptmann. Which thousand words are worth a picture? Experiments on video retrieval using a thousand concepts. In *ICME*, pages 41–44. IEEE, 2006.

[7] G. A. Miller. Wordnet: A lexical database for english. *Comm. ACM*, 38:39–41, 1995.

[8] 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*, 13(3):86–91, 2006.

[9] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *ACM Multimedia '07*, pages 991–1000, 2007.

[10] 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 '06*, pages 143–152, 2006.

[11] P. Over, T. Ianeva, W. Kraaij, and A. Smeaton. TRECVID 2005 - an overview. In *TRECVid*. NIST, USA, 2005.

[12] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, 1995.

[13] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC. In *TREC*, pages 21–30, 1992.

[14] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Comm. ACM*, 18(11):613–620, 1975.

[15] A. Schutz and T. Luckmann. *The structures of the life-world*. Heinemann, London, 1974.

[16] A. F. Smeaton. Techniques used and open challenges to the analysis, indexing and retrieval of digital video. *Inf. Syst.*, 32(4):545–559, 2007.

[17] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06*, pages 321–330. ACM Press, 2006.

[18] C. G. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM Multimedia '06*, pages 421–430. ACM, 2006.

[19] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Trans. Multimedia*, 9(5): 975–986, 2007.

[20] J. Tešić, A. P. Natsev, and J. R. Smith. Cluster-based data modeling for semantic video search. In *CIVR '07*, pages 595–602. ACM, 2007.

[21] M. F. Triola. *Essentials of Statistics*. Addison-Wesley, Boston, 2002.

[22] T. Volkmer, J. R. Smith, and A. P. Natsev. A web-based system for collaborative annotation of large image and video collections: an evaluation and user study. In *ACM Multimedia '05*, pages 892–901. ACM, 2005.

[23] D. Wang, X. Li, J. Li, and B. Zhang. The importance of query-concept-mapping for automatic video retrieval. In *ACM Multimedia '07*, pages 285–288. ACM, 2007.

[24] X.-Y. Wei and C.-W. Ngo. Ontology-enriched semantic space for video search. In *ACM Multimedia '07*, pages 981–990. ACM, 2007.

[25] M. Worring and G. Schreiber. Semantic image and video indexing in broad domains. *IEEE Trans. Multimedia*, 9(5), 2007.

[26] R. Yan and A. G. Hauptmann. A review of text and image retrieval approaches for broadcast news video. *Inf. Retr.*, 10 (4-5):445–484, 2007.