The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia

Cees G.M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W.M. Smeulders
ISLA, Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ, Amsterdam, The Netherlands
{cgmsnoek, worring, jvgemert, mark, smeulders}@science.uva.nl

ABSTRACT

We introduce the challenge problem for generic video indexing to gain insight in intermediate steps that affect performance of multimedia analysis methods, while at the same time fostering repeatability of experiments. To arrive at a challenge problem, we provide a general scheme for the systematic examination of automated concept detection methods, by decomposing the generic video indexing problem into 2 unimodal analysis experiments, 2 multimodal analysis experiments, and 1 combined analysis experiment. For each experiment, we evaluate generic video indexing performance on 85 hours of international broadcast news data, from the TRECVID 2005/2006 benchmark, using a lexicon of 101 semantic concepts. By establishing a minimum performance on each experiment, the challenge problem allows for component-based optimization of the generic indexing issue, while simultaneously offering other researchers a reference for comparison during indexing methodology development. To stimulate further investigations in intermediate analysis steps that influence video indexing performance, the challenge offers to the research community a manually annotated concept lexicon, pre-computed low-level multimedia features, trained classifier models, and five experiments together with baseline performance, which are all available at `http://www.mediamill.nl/challenge/`.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*

General Terms

Algorithms, Experimentation, Performance

Keywords

Video analysis, baseline, generic concept detection

MM'06, October 23–27, 2006, Santa Barbara, California, USA.

1. INTRODUCTION

The field of multimedia indexing has witnessed a rapid growth in recent years. Fueled by ever increasing capture, storage, and transmission capabilities, multimedia assets have become commonplace items to record, distribute, and share. We reached a point where users require instant access to their expanding repositories of multimedia data. Pushed by this demand, powerful multimedia analysis techniques have emerged. It has yielded a proliferation of methods, often evaluated on specific and small data sets. As a result, experiments are non-repeatable; making it hard to judge whether approaches are truly promising. Repeatable experiments, using published benchmarks, have been identified at the latest ACM SIGMM retreat as one of the requirements for the field to progress further [1].

In more mature fields, like computer vision, repeatable benchmark experiments have fostered the state-of-the-art. For problems as diverse as human gait analysis [2], color constancy [3], face recognition [4], and object detection [5,6] the availability of repeatable benchmark experiments has given researchers an environment to measure what factors affect performance most. Hence, it allows for an in-dept understanding of the problem at stake. In [2] for example, Sarkar *et al.* study gait-based identification of humans on a large data set. They decompose the problem into a number of components, for which they provide a standard implementation. The authors quantify the quality of each component by repeatable experiments on labeled data. By establishing a minimum performance on each part, the authors allow for component-based optimization of the problem, while at the same time offering other researchers a reference for comparison during methodology development. Hence, the authors make a transition from a benchmark to a *challenge problem.* From [2–6] it follows that a challenge problem requires a shared data set, a collection of repeatable experiments, a baseline implementation, and its performance.

1.1 Multimedia Indexing Challenge Problem

To arrive at a challenge problem for multimedia indexing, we first need shared multimedia data. A shared data set has always been a delicate issue. Multimedia archives are fragmented and mostly inaccessible due to copyrights and the sheer volume of data involved. Making it hard, often impossible even, for researchers world wide to share resources. As a consequence, comparison of systems has traditionally been difficult. To counter this trend, the American National Institute of Standards and Technology (NIST) initiated the

TREC Video Retrieval Evaluation (TRECVID) [7, 8]. The aim of the benchmark is to promote progress in content-based retrieval from digital video archives via open, metrics-based evaluation using a common large data set. Tasks include camera shot segmentation, camera motion detection, story segmentation, semantic concept detection, and several retrieval questions. The research community at large has joined this initiative. TRECVID has become the *de facto* data set to evaluate multimedia indexing research.

TRECVID has been of pivotal importance in assessing complete multimedia indexing methods on their relative merit. It has, however, not addressed the important issue of experiment repeatability of intermediate analysis steps on training data. This is mainly caused by the fact that TRECVID focuses on the final result of a multimedia processing system, be it a shot segmentation or a ranked list of fragments resulting from an interactive session with a video search engine. In theory, the TRECVID experiments are repeatable, but not on a system component level. Because TRECVID ignores intermediate results, component-based optimization and comparison during methodology development are impossible in practice.

Given the TRECVID data, what exactly is needed for a challenge problem in the multimedia field? To answer the question, we first focus on the fundamental problem in multimedia indexing that almost all research papers in the field address: the *semantic gap* [9]. This gap is defined as the discrepancy between machine computable low-level features on one end, and its semantic interpretation by humans on the other end. Since a large majority of work in multimedia research aims for bridging the semantic gap, see e.g. [10] for a bundled collection, the raised question can be rephrased as: what is needed to bridge the semantic gap?

Early approaches aiming to bridge the semantic gap focused on the feasibility of mapping low-level features, e.g. color, pitch, and term frequency, directly to high-level semantic concepts, like *commercials* [11], *nature* [12], and *baseball* [13]. This has yielded a variety of dedicated methods, which exploit simple decision rules to map low-level features to a single semantic concept. This specific detector approach will fail, however, when we aim for large-scale automated annotation of video archives. It is simply unfeasible to develop a tailor-made detector for every possible concept one can think of. Specific methods have aided in demonstrating the potential of semantic concept detection. For a challenge problem, however, we urge for an alternative.

Recently, generic approaches for concept detection [14–17] emerged as an adequate alternative for specific methods. Generic approaches learn a wide variety of concepts from a shared set of low-level features, often fused in various ways [16]. In contrast to specific methods, these approaches exploit the observation that mapping multimedia features to concepts requires quite many decision rules. To distill these rules, the methods make exhaustive use of machine learning. The machine learning paradigm has proven to be quite successful in terms of generic detection, as well as overall TRECVID benchmark performance. A challenge problem should aim for generic concept detection using machine learning.

Ideally, a generic video indexing system should learn and infer concepts from the multimedia data directly. However, the present day paradigm of choice in generic video indexing is to learn the concept classification rules by supervised learning. Supervised learning requires labeled examples. Hence, annotations are a valuable resource for generic concept detection. Moreover, when aiming for repeatability of experiments this ground truth needs to be shared. To cope with the demand for shared annotations in multimedia research, Lin *et al.* initiated a collaborative annotation effort in the TRECVID 2003 benchmark [18]. Guided by tools from Christel *et al.* [19] and Volkmer *et al.* [20] a common annotation effort was again started for the TRECVID 2005 benchmark. It has yielded a large and accurate set of labeled examples for a lexicon of 39 concepts, taken from a predefined concept ontology for multimedia [21, 22]. At present, efforts to produce a manually annotated lexicon of 1,000 concepts are underway [23]. Driven by the TRECVID benchmark various sets of annotated concepts have become publicly available.

A challenge problem is more than just manual annotations. In addition to concept examples, a challenge problem aiming to bridge the semantic gap by means of automatically detected high-level concepts requires intermediate results in the form of pre-computed low-level features, and a supervised learner. This offers fellow researchers the opportunity to focus on a single aspect of the generic video indexing problem, e.g. indexing based on visual analysis only, or a combined effort using fused versions of visual and textual analysis for example. In addition, researchers from pattern recognition or information retrieval can step in without the need to do expensive multimedia processing, since they can exploit the provided low-level features. It should be noted that in the course of the TRECVID benchmark some groups have donated features, most notably are the camera shot segmentation by CLIPS-IMAG [24], speech recognition results donated by LIMSI [25] and various multimedia features donated by Informedia [26]. In addition, all participants share their results on common test data for a limited lexicon of typically 10 high-level concepts. To date, however, nobody has provided low-level features and detected semantic concepts for a large lexicon on both training and test data, while these are crucial assets for any challenge problem.

Once a challenge problem is defined for multimedia indexing it allows to focus on new research frontiers. One of many open issues is to understand why a particular technique is suited best, or unsuited, for a specific class of semantic concepts. A multimedia indexing challenge problem lays the foundation for conceptual meta-analysis methods that investigate what strategy should be employed for a particular class of concepts.

1.2 Contribution

We describe in this paper the challenge problem for the automated detection of a lexicon of 101 semantic concepts in video. The purpose of the challenge problem is to gain insight in intermediate analysis steps that play a role in generic video indexing, by providing researchers with a framework for the systematic evaluation of video indexing components, while at the same time ensuring repeatability of experiments. To arrive at a challenge problem, we provide a general scheme for the systematic examination of automated concept detection methods, decomposing the generic video indexing problem into 2 unimodal analysis experiments, 2 multimodal analysis experiments, and 1 combined analysis experiment. For each experiment, we provide a baseline

implementation and its performance on TRECVID data. To stimulate further investigations in factors that influence generic video indexing performance, the challenge offers to the research community an annotated lexicon of 101 concepts, low-level multimedia features, trained classifier models, and baseline performance for the five experiments, which are available at `http://www.mediamill.nl/challenge/`.

The remainder of the paper is organized as follows. We first define the challenge problem in more detail. In section 3, we describe the baseline algorithm, which we exploit to learn 101 concepts in a generic fashion from low-level multimedia features. We present the evaluation with baseline performance and conceptual meta-analysis in section 4. We wrap up in the conclusions.

2. CHALLENGE PROBLEM DEFINITION

The purpose of the challenge problem for generic video indexing is to provide researchers with a framework for the systematic evaluation of video indexing components. To allow for systematic evaluation, we organize the challenge problem as a laboratory test [27]. In such a test the variability stemming from multimedia data, concepts, experiments, and performance must be structured to allow for comparison of results. To arrive at a laboratory test for the challenge problem, we separate a multimedia archive in a training set and a test set, using camera shots as the unit for indexing and evaluation, in line with the common procedure in literature [7, 8, 14–17]. For each set, we provide manually labeled ground truth, at the shot level, in the form of a shared concept lexicon. We define a set of experiments which index shots in the test set based on algorithms tuned on the training set. For each concept in the lexicon this should yield a list of shots, ranked according to detector confidence of concept presence. To evaluate these ranked lists we use standard measures from information retrieval. We will now describe the challenge problem in more detail.

2.1 Data Set

2.1.1 Multimedia Data

A publicly available archive of video data is a prerequisite for a challenge problem. In addition to this availability requirement, the archive of choice should provide a sufficiently large research challenge. To that end, a provocative video archive is, first of all, sizeable enough to allow for a diversity of experiments. Secondly, as it is meant for multimedia experiments, it should emphasize the multimedia nature of the data, i.e. containing speech in addition to the visual information. Thirdly, the videos should have a common granularity, e.g. camera shots, to provide a standardized basis for evaluation. The 2005 TRECVID corpus meets our demands.

The video archive of the 2005 TRECVID benchmark is composed of 169 hours of Arabic, Chinese, and US broadcast news sources, recorded in MPEG-1 during November 2004 by the Linguistic Data Consortium. The training data contains about 85 hours. The video archive comes together with automatic speech recognition results and machine translations donated by a US government contractor. Where it should be noted that both the speech recognition and machine translations yield noisy detection results. What is more, due to the machine translation, the text is unsynchronized with the visual content. Hence, the corpus provides a challenging basis for multimedia analysis. As an aside we note that the 2005 data will be reused in TRECVID 2006 together with new test data, assuring for researchers a broad applicability of developed algorithms. For all videos, the Fraunhofer Institute [28] provided a camera shot segmentation. Dublin City University created a common set of key frames [8]. The video data and key frames have been distributed to 57 teams from academic and corporate research labs, spread over 5 continents, already. The Linguistic Data Consortium aims to make the official release of the video data available for all interested parties soon [29]. The shot segmentation, automatic speech recognition results, and machine translations are available from NIST [30]. The 85 hours of training data from the TRECVID 2005 corpus forms the basis for the challenge problem. We divided this archive a priori into a non-overlapping train and test set. The challenge train set $\mathcal{A}$ contains 70% of the data, and the challenge test set $\mathcal{B}$ holds the remaining 30%. These sets form the basis for our lexicon of high-level concepts.

2.1.2 Annotated Concept Lexicon

Given the TRECVID corpus, we face the task of defining a lexicon of semantic concepts that our challenge problem should detect. Similar to [6], we choose concepts at random, but we take a predefined concept ontology for multimedia [22] as leading example. Concepts in this ontology are chosen based on extensive analysis of video archive query logs. Concepts should be related to program categories, setting, people, objects, activities, events, and graphics. In addition, a primary design choice was that concepts need to be clear by looking at a static key frame only. It has resulted in a lexicon of 39 concepts, which formed the basis for the TRECVID 2005 common annotation effort [20]. In part, we rely on this provided ground truth. We manually extend both the number of concepts and the number of annotations by browsing the shots in the training data, using our MediaMill video search engine [31]. To relieve the effort, we focus on positive instances of concepts adhering to the above categorization only. Presence of a concept was assumed to be binary, i.e. it is visible during a shot or not. Hence, the location of a concept in the image frame is not taken into account. Moreover, if the concept is true for some frame within the shot, then it was true for the entire shot. To assure a sound basis for supervised learning, concepts are added to the lexicon only when at least 30 positive instances are identified. To limit the need for disambiguation, only one person annotated the data. The manual annotation process has yielded an incomplete, but reliable ground truth for a lexicon of 101 semantic concepts, see Fig. 1 for visual examples. As new concepts and names keep appearing and disappearing in our world, these 101 concepts are bound to keep changing over time. However, by fixing the data set and concept lexicon, we allow for the systematic examination of automated concept detection methods. We provide statistics for the concept lexicon in overview Table 1 at the end of this paper.

2.2 Experiments

To arrive at a set of experiments for the automated indexing of 101 semantic concepts, we build on successful previous work in generic concept detection, e.g. [14–17]. Similar to this work, we perceive concept detection in video as a pattern recognition problem. Given pattern $\vec{x}$, part of a shot i,

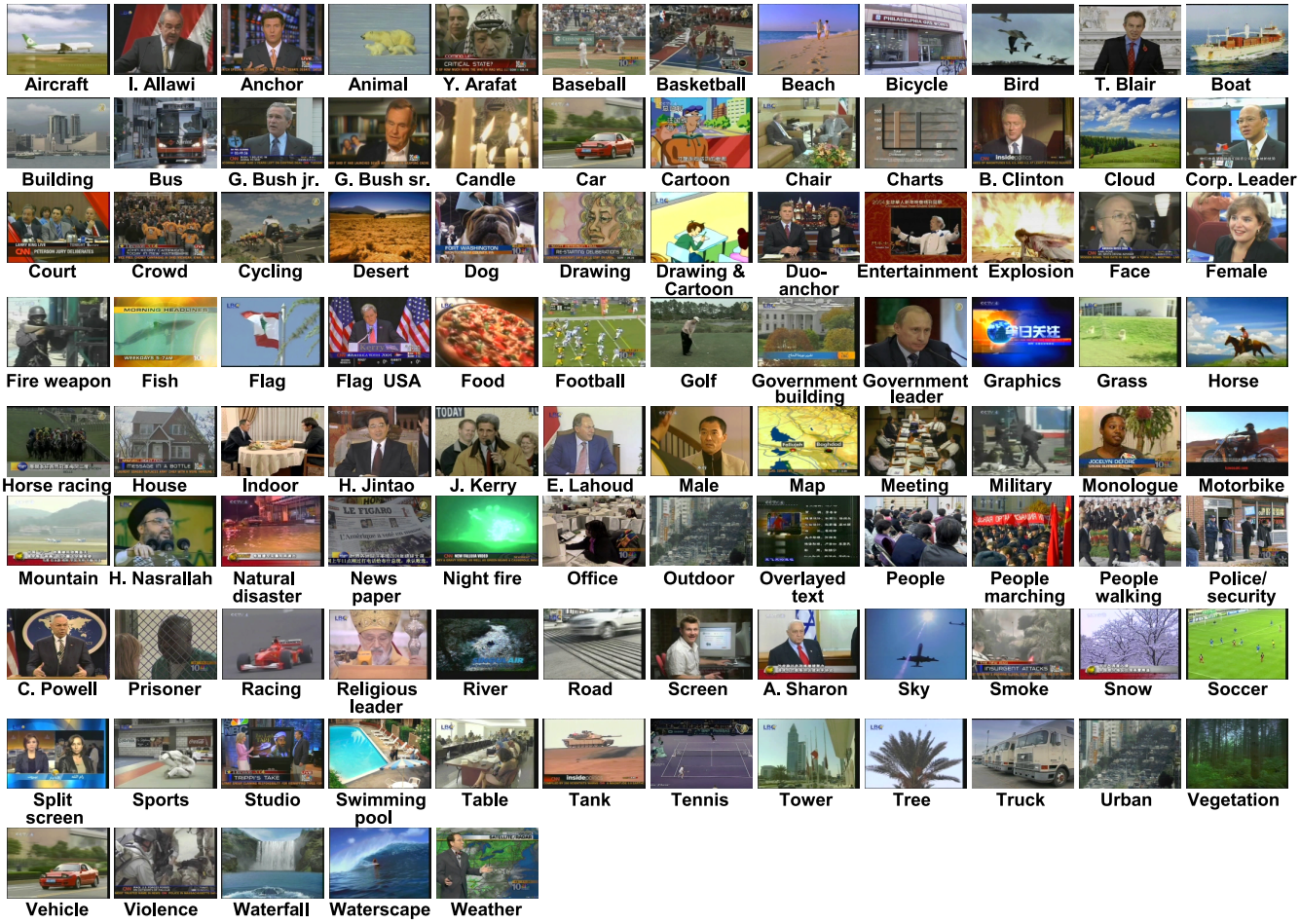

Figure 1: Visual impression of the 101 semantic concepts, which we detect within the challenge problem.

the aim is to obtain a probability measure, which indicates whether semantic concept ω_j is present in shot i. In pattern recognition, the strict definition of a probability depends on many factors and assumptions. Hence, it can not form the basis for comparison between different methods. Therefore, we do not use the probability directly. Instead, we utilize the probability as a confidence score, defined as $p(\omega_j|\vec{x}_i)$. To allow for metric-based evaluation, we employ ranking operator Φ to rank all shots based on the confidence score. This yields ranked list ρ_j, defined as:

$$\rho_j = \Phi\left(\{p(\omega_j|\vec{x}_i)\}_{i=1,2,\ldots,n}\right), \quad (1)$$

where n denotes the number of shots in the data set. Thus, each experiment uses supervised learning to convert a set of feature vectors into a ranked list of shots, ordered by concept detection confidence. The challenge experiments differ in the way they obtain feature vector $\vec{x}_i$.

In literature, the two most common approaches to acquire feature vector $\vec{x}_i$ from video are unimodal and multimodal content analysis. Considering unimodal analysis, we distinguish three data streams or modalities, namely the auditory modality, the textual modality, and the visual one. As speech is often the most informative part of the auditory source, the challenge experiments focus on textual features obtained from transcribed speech and on visual features obtained from key frames. For multimodal content analysis the visual and textual streams need to be fused at some point. We consider two classes of fusion schemes, namely early fusion and late fusion [16]. Naturally, the above approaches for generic video indexing may be combined. In fact, previous work [16, 17] indicates that the optimal analysis often varies per concept. The challenge experiments address 2 unimodal, 2 multimodal, and 1 combined analysis approach. We sketch the data flow for all five experiments in Fig. 2.

In the first challenge experiment we focus on a pure visual analysis of multimedia data. The challenge is to learn semantic concepts from a visual feature vector $\vec{v}_i$. Despite a wide variety of visual analysis methods proposed in literature [9], there is no consensus yet on what visual feature representation to choose for effective generic concept detection. We therefore identify the following experiment:

- **Experiment 1:** Given a visual feature vector, $\vec{v}_i$, learn for each of the 101 semantic concepts ω_j a ranked list ρ_j^1;

In contrast to visual analysis, textual analysis is a well understood problem. Standard techniques have proven to be useful in a video indexing setting also, even when the text feature vector $\vec{t}_i$ results from noisy speech recognition [26].

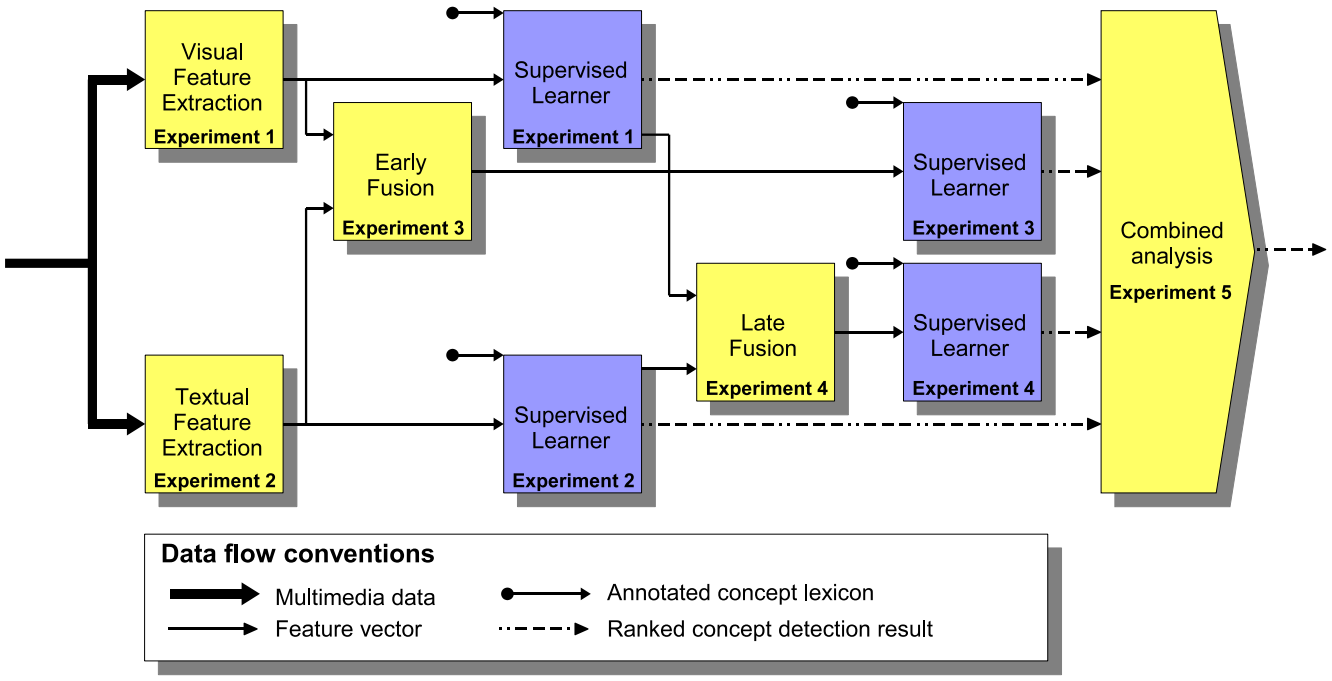

Figure 2: Data flow within the proposed challenge problem for generic video indexing of 101 semantic concepts. Experiment 1 and 2 focus on unimodal analysis, yielding a visual and a textual concept classification. Experiment 3 and 4 employ an early and late fusion scheme respectively. The challenge problem allows for the construction of four classifiers for each concept. In experiment 5, an optimum is selected based on combined analysis.

Recall that apart from muddled transcripts, the text from the TRECVID data also suffers from unsynchronized and noisy machine translations. Under such heavy circumstances, coping with textual data offers quite a challenge indeed. We identify the following experiment:

- **Experiment 2:** Given a textual feature vector, $\vec{t}_i$, learn for each of the 101 semantic concepts ω_j a ranked list ρ_j^2;

Indexing approaches that rely on early fusion first extract unimodal features. After analysis of the various unimodal streams, the extracted features are combined into a multimodal feature representation $\vec{e}_i$. Subsequently, early fusion methods rely on supervised learning to classify semantic concepts. Early fusion yields a truly multimedia feature representation, since the features are integrated from the start. Disadvantage of the approach is the difficulty to combine features into a common representation. Moreover, early fusion suffers from features with poor quality We identify the following challenge experiment:

- **Experiment 3:** Given an early fusion feature vector, $\vec{e}_i$, learn for each of the 101 semantic concepts ω_j a ranked list ρ_j^3;

Late fusion approaches also start with extraction of unimodal features. In contrast to early fusion, where features are then combined into a multimodal representation, approaches for late fusion learn semantic concepts directly from unimodal features. Hence, the dimensionality of the problem is reduced with the potential of easier analysis. In general, late fusion schemes combine learned unimodal concept detection scores into a multimodal representation $\vec{l}_i$. Then late fusion methods rely on supervised learning to classify semantic concepts. Late fusion focuses on the individual strength of modalities. Unimodal concept detection scores are fused into a multimodal semantic representation rather than a feature representation. A disadvantage of late fusion schemes is their expensiveness in terms of the learning effort, as every modality requires a separate supervised learning stage. Moreover, the combined representation requires an additional learning stage. We identify:

- **Experiment 4:** Given a late fusion feature vector, $\vec{l}_i$, learn for each of the 101 semantic concepts ω_j a ranked list ρ_j^4;

Given the large variety in semantic concepts, it is unlikely that each concept requires a similar analysis approach. A *tree* for example, is best detected in the visual content. In contrast, *Tony Blair* at the current level of person recognition is almost exclusively detectable using text. We identify a combined analysis experiment to gain insight in the role of various analysis approaches on concept detection performance. Based on the ranked lists from the previous four challenge experiments, various combined analysis methods can be defined, which ultimately yield an optimum combined ranked list. Hence, given the previous four experiments, we identify the final experiment:

- **Experiment 5:** Given the four ranked lists, ρ_j^1, ρ_j^2, ρ_j^3, ρ_j^4, from the previous four experiments, learn for each of the 101 semantic concepts ω_j an optimum combined ranked list, ρ_j^5;

2.3 Performance Metric

We use *average precision* to determine the accuracy of ranked concept detection results on our experiments, following the standard in TRECVID evaluations. The average precision is a single-valued measure that is proportional to the area under a recall-precision curve. This value is the average of the precision over all relevant judged shots. Hence, it combines precision and recall into one performance value. Let $\rho^k = \{i_1, i_2, \ldots, i_k\}$ be a ranked version of the answer set A. At any given rank k let $R \cap \rho^k$ be the number of relevant shots in the top k of ρ, where R is the total number of relevant shots. Then average precision, AP, is defined as:

$$AP(\rho) = \frac{1}{R} \sum_{k=1}^{A} \frac{R \cap \rho^k}{k} \psi(l_k) \ , \qquad (2)$$

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

3. BASELINE IMPLEMENTATION

We provide a baseline implementation for each experiment using standard algorithms from literature. By establishing a minimum performance on each experiment, the challenge problem allows for component-based optimization of the generic indexing issue, while at the same time offering other researchers a reference for comparison during indexing methodology development. Note that researchers may compare against the entire system or its components. The implementation of our baseline algorithm is structured according to the data flow sketched in Fig. 2. We will now briefly explain the components of the algorithm.

3.1 Supervised Learner

We choose from a large variety of supervised machine learning approaches to obtain confidence measure $p(\omega_j|\vec{x}_i)$. The Support Vector Machine (SVM) framework [32] has proven to be a solid choice [15–17]. Here we use the LIBSVM implementation [33] with radial basis function. The usual SVM method provides a margin in the result. We prefer Platt's conversion method [34] to achieve a confidence score. SVM classifiers thus trained for ω_j, result in an estimate $p(\omega_j|\vec{x}_i, \vec{q})$, where $\vec{q}$ are parameters of the SVM. The influence of the SVM parameters on concept detection is significant [35]. We obtain good parameter settings by using an iterative search on a large number of SVM parameter combinations. We measure average precision performance of all parameter combinations and select the combination that yields the best performance, $\vec{q}^*$. Here we use a 3-fold cross validation on train set $\mathcal{A}$ to prevent overfitting of parameters. The result of the parameter search over $\vec{q}$ is the improved model $p(\omega_j|\vec{x}_i, \vec{q}^*)$, contracted to $p^*(\omega_j|\vec{x}_i)$.

3.2 Visual Feature Extraction

Visual feature extraction is based on the method described in [36]. In short, the procedure first extracts a number of color invariant texture features per pixel. Based on these features, it labels a set of predefined regions in a key frame image with similarity scores for a total of 15 low-level visual concepts, like *road*, *sky*, *water body*, and so on. This yields a 15-bins histogram, where each bin represents a similarity score to one of the 15 regional concepts. 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, $\vec{v}_i$. To learn semantic concepts, $\vec{v}_i$ serves as the input for the supervised learner.

3.3 Textual Feature Extraction

In the textual modality, we learn the association between transcribed speech and concepts, see [17]. We map the Chinese and Arabic story level machine translations to shot level using linear interpolation. To learn the relation between uttered speech and concepts, we connect stemmed and stopped words to shots. We make this connection within the temporal boundaries of a shot. We derive a vocabulary of uttered words that co-occur with concept ω_j using the shot-based annotations of the training data. For each concept ω_j, we learn a separate vocabulary, Λ^{ω_j}, as the uttered words are specific for that concept. Since a news anchor or reporter often mentions indicative words just before or after a concept is visible, we stretch the shot boundaries by inclusion of the previous and next shot on each side. For feature extraction we compare the text associated with the stretched shot with Λ^{ω_j}. This comparison yields a text vector $\vec{t}_i$ for shot i, which contains the histogram of the words in association with ω_j. To learn semantic concepts, $\vec{t}_i$ serves as the input for the supervised learner.

3.4 Early Fusion

For the early fusion experiment, we combine the feature vectors resulting from visual feature extraction and textual feature extraction. We adopt the method proposed in [16], using vector concatenation to unite the features $\vec{v}_i$ and $\vec{t}_i$. After feature normalization, we obtain early fusion vector $\vec{e}_i$. To learn semantic concepts, $\vec{e}_i$ serves as the input for the supervised learner.

3.5 Late Fusion

We again follow [16] for the late fusion experiment. Recall that late fusion requires two supervised learning stages. We consider the size of the used sets an implementation issue. Therefore, we split train set $\mathcal{A}$ into two sets: $\mathcal{A}^1$ and $\mathcal{A}^2$, each containing 50% of the data. We utilize set $\mathcal{A}^1$ to obtain a confidence score after visual analysis, i.e. $p^*(\omega_j|\vec{v}_i)$, and a confidence score resulting from textual analysis, i.e. $p^*(\omega_j|\vec{t}_i)$. We concatenate $p^*(\omega_j|\vec{v}_i)$ with $p^*(\omega_j|\vec{t}_i)$, into late fusion vector $\vec{l}_i$. Then $\vec{l}_i$ serves as the input for the supervised learner, which learns semantic concepts on set $\mathcal{A}^2$.

3.6 Combined Analysis

Each of the four previous experiments results in an optimized ranking per concept. We measure average precision performance according to 3-fold cross validation, for each concept and each experiment, on set $\mathcal{A}$. Similar to [17], we select per concept the experiment that maximizes performance on training data:

$$\rho_j^5 = \max\left(AP(\rho_j^1), AP(\rho_j^2), AP(\rho_j^3), AP(\rho_j^4)\right). \qquad (3)$$

4. BASELINE PERFORMANCE

We establish a baseline performance for each of the five challenge experiments using the baseline algorithm. The

Table 1: Overview of the the challenge problem for automated concept detection in multimedia, showing 101 concepts and the percentage of positively labeled examples used for the training set and the test set, together with average precision results for the five challenge experiments on test data. Concepts are ordered based on the training samples used for learning.

		Ground Truth		Challenge Experiments							Ground Truth		Challenge Experiments				
	Concept	Train (%)	Test (%)	1	2	3	4	5		Concept	Train (%)	Test (%)	1	2	3	4	5
1	People	77.67	75.87	0.831	0.817	0.890	0.840	0.840	52	Table	0.75	0.52	0.073	0.006	0.037	0.060	0.073
2	Face	64.15	62.37	0.895	0.737	0.892	0.890	0.890	53	Tower	0.75	0.63	0.057	0.009	0.023	0.033	0.057
3	Overlayed text	36.33	34.30	0.669	0.533	0.642	0.666	0.669	54	Basketball	0.69	0.34	0.382	0.219	0.179	0.239	0.382
4	Outdoor	32.68	38.33	0.688	0.579	0.709	0.691	0.691	55	Y. Arafat	0.62	0.88	0.026	0.072	0.034	0.013	0.072
5	Entertainment	19.64	12.55	0.166	0.179	0.257	0.146	0.179	56	Chair	0.60	0.58	0.486	0.101	0.261	0.467	0.486
6	Indoor	19.59	21.20	0.593	0.460	0.592	0.606	0.592	57	Explosion	0.53	1.04	0.098	0.038	0.078	0.046	0.038
7	Studio	13.66	14.20	0.636	0.490	0.664	0.651	0.664	58	Food	0.50	0.83	0.287	0.085	0.188	0.170	0.287
8	People walking	13.61	16.83	0.353	0.294	0.338	0.296	0.338	59	Bus	0.43	0.64	0.013	0.007	0.009	0.005	0.007
9	Urban	11.78	8.80	0.222	0.178	0.195	0.201	0.195	60	Snow	0.41	0.53	0.085	0.018	0.045	0.004	0.085
10	Crowd	11.48	16.12	0.480	0.288	0.490	0.440	0.490	61	Fire weapon	0.35	0.52	0.121	0.013	0.060	0.047	0.121
11	Sky	10.77	11.38	0.478	0.218	0.496	0.463	0.496	62	Tennis	0.34	0.56	0.448	0.195	0.299	0.397	0.397
12	Government leader	9.35	7.87	0.213	0.213	0.222	0.236	0.213	63	Prisoner	0.33	0.22	0.047	0.027	0.051	0.004	0.051
13	Violence	8.07	9.75	0.317	0.301	0.334	0.237	0.334	64	News paper	0.31	0.27	0.375	0.000	0.121	0.384	0.375
14	Road	7.76	6.60	0.195	0.138	0.212	0.188	0.195	65	E. Lahoud	0.30	0.15	0.289	0.080	0.115	0.196	0.289
15	Vehicle	7.61	8.53	0.221	0.167	0.271	0.190	0.271	66	J. Kerry	0.29	0.01	0.000	0.012	0.002	0.001	0.000
16	Building	6.86	11.16	0.316	0.154	0.233	0.291	0.154	67	House	0.29	0.36	0.023	0.004	0.007	0.009	0.004
17	Male	5.71	2.38	0.086	0.034	0.068	0.069	0.086	68	Government building	0.27	0.19	0.011	0.038	0.079	0.002	0.002
18	Anchor	5.09	4.85	0.631	0.201	0.620	0.618	0.631	69	Religious leader	0.27	0.23	0.043	0.026	0.035	0.041	0.026
19	Car	4.87	5.93	0.252	0.118	0.246	0.215	0.252	70	Fish	0.27	0.12	0.489	0.068	0.408	0.312	0.489
20	Meeting	4.53	4.86	0.257	0.158	0.211	0.257	0.257	71	Duo-anchor	0.26	0.18	0.634	0.022	0.108	0.287	0.634
21	Female	4.38	2.11	0.086	0.020	0.061	0.068	0.086	72	Golf	0.25	0.31	0.091	0.007	0.042	0.143	0.091
22	Military	4.14	6.58	0.217	0.206	0.235	0.203	0.235	73	I. Allawi	0.21	0.02	0.000	0.030	0.002	0.000	0.000
23	Vegetation	3.87	4.64	0.183	0.051	0.161	0.150	0.183	74	Bicycle	0.20	0.04	0.006	0.454	0.223	0.733	0.454
24	Sports	3.76	2.61	0.304	0.267	0.231	0.308	0.304	75	Court	0.20	0.30	0.093	0.041	0.030	0.052	0.041
25	Monologue	3.10	2.33	0.094	0.051	0.074	0.081	0.094	76	G. Bush sr.	0.20	0.01	0.000	0.000	0.000	0.000	0.000
26	Graphics	2.89	3.48	0.365	0.275	0.379	0.367	0.365	77	Football	0.20	0.39	0.048	0.016	0.020	0.043	0.016
27	Corporate leader	2.57	1.30	0.016	0.020	0.014	0.018	0.020	78	Cycling	0.18	0.03	0.042	0.950	0.888	0.608	0.950
28	Waterscape	2.31	1.89	0.150	0.079	0.134	0.142	0.134	79	Bird	0.18	0.23	0.724	0.577	0.761	0.743	0.724
29	People marching	1.93	4.13	0.228	0.087	0.267	0.109	0.267	80	Drawing & Cartoon	0.17	0.38	0.265	0.207	0.181	0.191	0.207
30	Soccer	1.67	0.29	0.503	0.000	0.079	0.372	0.503	81	Horse	0.16	0.02	0.000	0.000	0.000	0.000	0.000
31	Mountain	1.64	1.01	0.141	0.022	0.092	0.157	0.141	82	Dog	0.14	0.38	0.225	0.012	0.103	0.019	0.012
32	G. Bush jr.	1.61	0.54	0.062	0.065	0.040	0.060	0.062	83	Night fire	0.14	0.05	0.526	0.001	0.249	0.000	0.526
33	Office	1.56	1.75	0.077	0.024	0.045	0.037	0.024	84	Horse racing	0.12	0.02	0.000	0.000	0.000	0.000	0.000
34	Screen	1.53	1.90	0.101	0.063	0.058	0.121	0.063	85	River	0.10	0.09	0.310	0.710	0.654	0.098	0.710
35	Flag	1.26	1.12	0.189	0.029	0.120	0.166	0.189	86	Racing	0.09	0.12	0.029	0.176	0.175	0.004	0.176
36	Truck	1.16	1.02	0.038	0.019	0.042	0.038	0.019	87	Candle	0.08	0.10	0.011	0.057	0.080	0.001	0.057
37	Map	1.16	1.21	0.476	0.220	0.313	0.407	0.476	88	Cartoon	0.08	0.21	0.259	0.671	0.278	0.285	0.671
38	Smoke	1.13	2.14	0.250	0.103	0.366	0.149	0.250	89	Drawing	0.08	0.17	0.293	0.011	0.044	0.026	0.293
39	Animal	1.00	0.91	0.209	0.204	0.199	0.239	0.199	90	Tank	0.08	0.08	0.008	0.003	0.011	0.004	0.008
40	Weather	0.99	1.25	0.405	0.730	0.701	0.566	0.730	91	Swimming pool	0.08	0.10	0.003	0.001	0.001	0.002	0.001
41	Aircraft	0.99	0.94	0.073	0.033	0.115	0.030	0.115	92	Beach	0.08	0.06	0.027	0.001	0.065	0.007	0.065
42	Police/security	0.92	0.77	0.012	0.053	0.082	0.017	0.053	93	Waterfall	0.07	0.08	0.381	0.011	0.415	0.001	0.011
43	Flag USA	0.92	0.94	0.227	0.036	0.157	0.184	0.227	94	Motorbike	0.05	0.16	0.006	0.029	0.007	0.005	0.006
44	Grass	0.90	0.59	0.064	0.004	0.028	0.054	0.064	95	T. Blair	0.05	0.26	0.005	0.031	0.015	0.048	0.048
45	Cloud	0.87	1.54	0.117	0.042	0.078	0.129	0.117	96	B. Clinton	0.05	0.21	0.004	0.010	0.189	0.002	0.189
46	Split screen	0.86	0.60	0.630	0.100	0.321	0.566	0.630	97	H. Nasrallah	0.05	0.19	0.006	0.068	0.004	0.001	0.006
47	Desert	0.81	1.44	0.103	0.032	0.093	0.052	0.032	98	C. Powell	0.05	0.47	0.010	0.022	0.085	0.008	0.010
48	Natural disaster	0.81	0.93	0.055	0.091	0.139	0.084	0.091	99	A. Sharon	0.04	0.19	0.050	0.019	0.035	0.001	0.050
49	Boat	0.78	0.54	0.096	0.109	0.083	0.020	0.109	100	H. Jintao	0.03	1.03	0.030	0.023	0.044	0.018	0.023
50	Tree	0.78	0.84	0.124	0.011	0.063	0.087	0.124	101	Baseball	0.01	0.41	0.003	0.066	0.003	0.011	0.003
51	Charts	0.76	0.51	0.327	0.301	0.254	0.355	0.327		Mean			0.216	0.147	0.201	0.191	0.237

baseline performance serves to illustrate the minimum result that is expected from any unimodal, multimodal, or combined video analysis method. For each experiment, we report the average precision per concept on test set $\mathcal{B}$ in Table 1.

4.1 Experiment Results

The baseline indicates that for 45 out of 101 concepts a visual only analysis with experiment 1 yields the best performance. Visual analysis is especially effective for concepts that often appear in uniform settings, e.g. sports like *tennis*, *basketball*, and *soccer*, or studio setting related concepts such as *anchor*, *split screen*, and *duo-anchor*. The baseline implementation for experiment 1 performs moderate for sparse concepts such as *candle* and *beach*. Learning from few examples is a general problem, however, which negatively influences all challenge experiments.

The text-based analysis in experiment 2 yields the best performance for 14 concepts only. This is not surprising as the text resulting from speech recognition and machine translations is of disputable quality. Text analysis does work for concepts that are transcribed with a specific and limited vocabulary. In such cases as *weather*, detection is therefore relatively easy based on textual content only. A textual analysis is often the best guess for sparse concepts, e.g. *baseball*, *Hassan Nasrallah*, and *motorbike*. In these cases a single word, e.g. a persons name, can be an important distinguishing feature. Note, however, that the difference with the other experiments is marginal.

Early fusion in experiment 3 obtains the best performance for 28 concepts. Early fusion works particularly well for concepts that have many positively labeled training samples, like *people*, *outdoor*, and *studio*. When both visual and textual analysis perform well in isolation, their early fusion

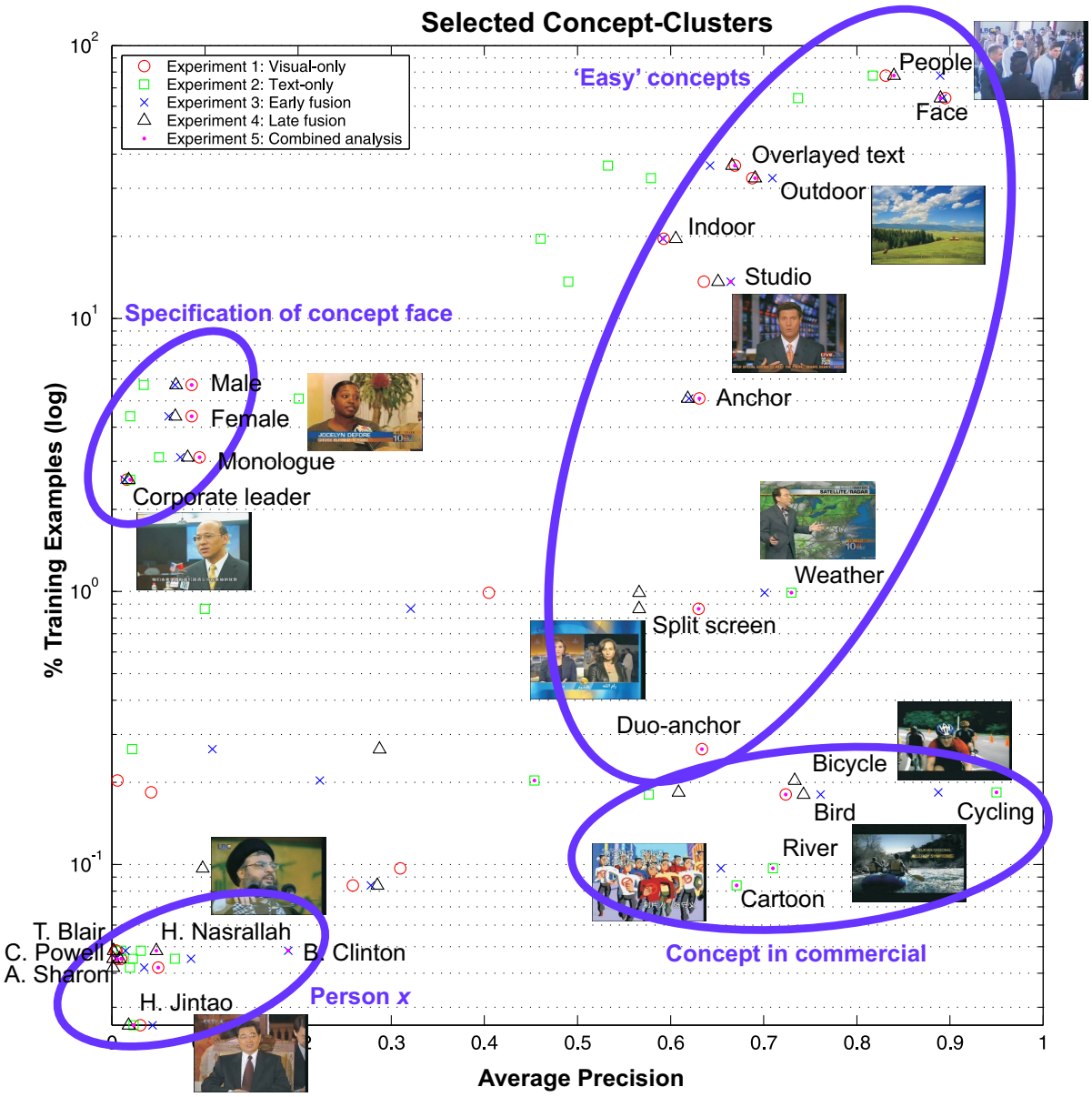

Figure 3: Selected concept-clusters that require special attention. Performance for 'easy' concepts needs to be raised to 1.0. Average precision of face-related concepts lacks behind, given the number of available training examples. Person x is problematic still. When concepts appear in commercials it could result in a misleading indication of indexing performance.

combination often yields good results also. Apparently, for concepts like *people marching*, *military*, and *natural disaster* the visual and textual features complement each other. In contrast, when one of the modalities yields bad indexing performance, due to poor quality text features for example, the combination may suffer. This is especially hurting the early fusion performance for concept *soccer*.

In experiment 4, late fusion is the best performer for 14 concepts. Similar to early fusion, video indexing using late fusion performs well when both modalities yield reasonable performance in isolation, e.g. *indoor*. In contrast to early fusion, late fusion is able to account for bad unimodal concept detection results by exploiting its first learning stage. For concepts such as *mountain*, *screen*, and *news paper*, late fusion learns to reduce the influence from the weak performing textual modality, yielding an optimal late fusion result. However, this is not always the case, indicating existence of a trade off between unimodal analysis performance and the number of examples used for training in the two learning stages of late fusion. We conclude from these results that an additional learning stage doesn't necessarily have a positive effect on performance.

The baseline implementation of combined analysis experiment 5 selects the best indexing approach for 50 out of 101 concepts. When we take the mean of the average precision over 101 concepts, this experiment yields the best overall

result. However, for more than half of the concepts the cross-validation performance on the training set is not the optimal estimator for test set performance. This indicates that much is to be gained when researchers employ more advanced techniques for the combination method.

4.2 Conceptual Meta-Analysis

The results in Table 1 provide ample opportunity for conceptual meta-analysis. We restrict ourselves here to four clusters of concepts that, in our view, require special attention. In Fig. 3, we highlight 25 selected concepts, clustered according to the number of training samples used and their average precision performance.

In general, the number of training samples has a positive effect on concept detection performance for all experiments. When the annotated samples include more than 5% of the training data it almost always results in a reasonable performance. For concepts such as *face*, *outdoor*, and *weather*, performance has even reached a robust level already. It is not a coincidence that these concepts appear often in evaluations of video indexing methods. A grand challenge for frequently appearing concepts is to raise the average precision performance towards 1.0, to allow for practical utility of video indexing technology in applications where (almost) perfect performance is required

In contrast to the concept *face*, semantically related concepts such as *female* and *monologue* perform quite bad still. This is surprising given the relative large amount of training examples available. It might indicate that visual and textual features are not the most discriminative features for these concept classes. In contrast, features related to the characteristics of the human voice, or features related to the recording circumstances might be better suited. More research is needed to accurately classify face-related concepts based on visual and/or textual features.

A special class of concepts is *person x*, i.e. named persons. A *person x* index is useful for video retrieval applications, but their detection is currently problematic. This is caused by sparseness and the high variability in the visual modality. Our experiments indicate that for the baseline a text-based analysis yields the most successful approach. In general, however, performance is disappointing for all baseline experiments. An obvious improvement would be the inclusion of face recognition techniques.

When a concept appears in a commercial, it may result in a misleading indication of performance. In such cases as *river* and *cycling*, performance is quite good based on a relatively small number of training examples. When we analyze results the reason is easily resolved: these concepts appear in commercials. In this case indexing boils down to (near) copy detection. Obviously, this is not what generic video indexing methods should aim for. How to handle commercials is an open issue in multimedia indexing research that needs to be dealt with as a separate problem.

5. CONCLUSIONS

In this paper, we present the challenge problem for automatic indexing of 101 semantic concepts in video. The challenge problem provides multimedia researchers with an experimental environment to measure the influence of individual video indexing system components and their combined usage. We identify five challenge experiments, by decomposing the generic video indexing problem into a visual-only, textual-only, early fusion, late fusion, and combined analysis experiment. We provide a baseline implementation for each experiment together with baseline results. By establishing a minimum performance on each experiment (Table 1), the challenge problem allows for component-based optimization of the generic indexing issue, while simultaneously offering other researchers a reference for comparison during indexing methodology development. Hence, it allows to gain insight in factors that affect performance of multimedia analysis methods, while at the same time fostering repeatability of experiments.

The challenge offers to the research community a manually annotated lexicon containing 101 semantic concepts, pre-computed low-level multimedia features, trained classifier models, and baseline experiment performance for five pre-cooked experiments on 85 hours of publicly available TRECVID 2005 video data. Fellow multimedia indexing researchers may use the challenge problem by replacing one or more components of the baseline implementation (Fig. 2) for one or more of their own algorithms. In addition, the baseline concept detection can be a valuable resource for (interactive) video retrieval experiments. We anticipate that the availability of the challenge problem will greatly facilitate the reliable evaluation of generic multimedia indexing algorithms, and make it easier for researchers in the multimedia indexing field to compare their algorithms. Furthermore, our challenge lowers the threshold for researchers from other disciplines to enter the field of multimedia analysis.

6. ACKNOWLEDGMENTS

This research is sponsored by the BSIK MultimediaN project. The authors are grateful to NIST and the TRECVID coordinators for the benchmark organization effort. We thank the TRECVID community for the 2005 common annotation effort.

7. REFERENCES

[1] L.A. Rowe and R. Jain. ACM SIGMM retreat report on future directions in multimedia research. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 1(1):3–13, 2005.

[2] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, and K.W. Bowyer. The humanID gait challenge problem: Data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[3] K. Barnard, L. Martin, B. Funt, and A. Coath. A data set for color research. *Color Research & Application*, 27(3):147–151, 2002.

[4] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE International Conference on Computer Vision and Pattern Recognition*, San Diego, USA, 2005.

[5] M. Everingham et al. The 2005 pascal visual object classes challenge. In *Selected Proceedings of the First PASCAL Challenges Workshop*, LNAI. 2006.

[6] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

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

[8] A.F. Smeaton. Large scale evaluations of multimedia information retrieval: The TRECVid experience. In *CIVR*, volume 3569 of *LNCS*, pages 19–27. Springer-Verlag, 2005.

[9] 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 Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[10] H.J. Zhang, J.R. Smith, and Q. Tian, editors. *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*. Singapore, 2005.

[11] R. Lienhart, C. Kuhmünch, and W. Effelsberg. On the detection and recognition of television commercials. In *IEEE Conference on Multimedia Computing and Systems*, pages 509–516, Ottawa, Canada, 1997.

[12] J.R. Smith and S.-F. Chang. Visually searching the web for content. *IEEE Multimedia*, 4(3):12–20, 1997.

[13] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for TV baseball programs. In *ACM Multimedia*, pages 105–115, Los Angeles, USA, 2000.

[14] M.R. Naphade and T.S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151, 2001.

[15] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M.R. Naphade, A.P. Natsev, C. Neti, H.J. Nock, J.R. Smith, B.L. Tseng, Y. Wu, and D. Zhang. IBM research TRECVID-2003 video retrieval system. In *Proceedings of the TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2003.

[16] C.G.M. Snoek, M. Worring, and A.W.M. Smeulders. Early versus late fusion in semantic video analysis. In *ACM Multimedia*, pages 399–402, Singapore, 2005.

[17] 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 Transactions on Pattern Analysis and Machine Intelligence*, 28(10), 2006.

[18] C.-Y. Lin, B.L. Tseng, and J.R. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *Proceedings of the TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2003.

[19] M. Christel, T. Kanade, M. Mauldin, R. Reddy, M. Sirbu, S. Stevens, and H. Wactlar. Informedia digital video library. *Communicationns of the ACM*, 38(4):57–58, 1995.

[20] T. Volkmer, J.R. Smith, A.P. Natsev, M. Campbell, and M. Naphade. A web-based system for collaborative annotation of large image and video collections. In *ACM Multimedia*, Singapore, 2005.

[21] A.G. Hauptmann. Towards a large scale concept ontology for broadcast video. In *International Conference on Image and Video Retrieval*, volume 3115 of *LNCS*, pages 674–675. Springer-Verlag, 2004.

[22] M.R. Naphade, L. Kennedy, J.R. Kender, S.-F. Chang, J.R. Smith, P. Over, and A. Hauptmann. A light scale concept ontology for multimedia understanding for trecvid 2005. Technical Report RC23612, IBM T.J. Watson Research Center, 2005.

[23] M. Naphade, J.R. 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.

[24] G.M. Quénot, D. Moraru, L. Besacier, and P. Mulhem. CLIPS at TREC-11: Experiments in video retrieval. In E.M. Voorhees and L.P. Buckland, editors, *Proceedings of the 11th Text REtrieval Conference*, volume 500-251 of *NIST Special Publication*, Gaithersburg, USA, 2002.

[25] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1–2):89–108, 2002.

[26] H.D. Wactlar, M.G. Christel, Y. Gong, and A.G. Hauptmann. Lessons learned from building a terabyte digital video library. *IEEE Computer*, 32(2):66–73, 1999.

[27] J. Tague-Sutcliffe. The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, 28(4):467–490, 1992.

[28] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In *Proceedings of the TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2004.

[29] K. Walker. Linguistic data consortium, `http://www.ldc.upenn.edu/`, April 2006. Personal communication.

[30] P. Over. Trecvid data availability website, April 2006. `http://www-nlpir.nist.gov/projects/trecvid/trecvid.data.html/`.

[31] C.G.M. Snoek, M. Worring, J.C. van Gemert, J.M. Geusebroek, D.C. Koelma, G.P. Nguyen, O. de Rooij, and F.J. Seinstra. MediaMill: Exploring news video archives based on learned semantics. In *Proc. ACM Multimedia*, pages 225–226, Singapore, 2005.

[32] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, USA, 2nd edition, 2000.

[33] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

[34] J.C. Platt. Probabilities for SV machines. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 2000.

[35] M.R. Naphade. On supervision and statistical learning for semantic multimedia analysis. *Journal of Visual Communication and Image Representation*, 15(3):348–369, 2004.

[36] J.C. van Gemert, J.M. Geusebroek, C.J. Veenman, C.G.M. Snoek, and A.W.M. Smeulders. Robust scene categorization by learning image statistics in context. In *Int'l Workshop on Semantic Learning Applications in Multimedia, in conjunction with CVPR'06*, New York, USA, 2006.