High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements

Alan F. Smeaton[1] and Paul Over[2] and Wessel Kraaij[3]

[1] Dublin City University Alan.Smeaton@DCU.ie
[2] National Institute of Standards and Technology over@nist.gov
[3] TNO Wessel.Kraaij@tno.nl

Summary. *Successful and effective content-based access to digital video requires fast, accurate and scalable methods to determine the video content automatically. A variety of contemporary approaches to this rely on text taken from speech within the video, or on matching one video frame against others using low-level characteristics like colour, texture, or shapes, or on determining and matching objects appearing within the video. Possibly the most important technique, however, is one which determines the presence or absence of a high-level or semantic feature, within a video clip or shot. By utilizing dozens, hundreds or even thousands of such semantic features we can support many kinds of content-based video navigation. Critically however, this depends on being able to determine whether each feature is or is not present in a video clip. The last 5 years have seen much progress in the development of techniques to determine the presence of semantic features within video. This progress can be tracked in the annual TRECVid benchmarking activity where dozens of research groups measure the effectiveness of their techniques on common data and using an open, metrics-based approach. In this chapter we summarise the work done on the TRECVid high-level feature task, showing the progress made year-on-year. This provides a fairly comprehensive statement on where the state-of-the-art is regarding this important task, not just for one research group or for one approach, but across the spectrum. We then use this past and on-going work as a basis for highlighting the trends that are emerging in this area, and the questions which remain to be addressed before we can achieve large-scale, fast and reliable high-level feature detection on video. [4]

1 Introduction

Searching for relevant video fragments in a large collection of video clips is a much harder task than searching textual collections. A user's information need is more easily represented as a textual description in natural language using high-level concepts that directly relate to

[4] Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

the user's ontology which relates terminology to real world objects and events. Even though raw video clips lack textual descriptions, low-level signal processing techniques can however describe them in terms of color histograms, textures etc. The fact that there exists a mismatch between the low-level interpretation of video frames and the representation of an information need as expressed by a user is called the "semantic gap" [20].

Up to this point in time, video archives have overcome the semantic gap and can facilitate search by manual indexing of video productions, which is a very costly approach. The metadata produced this way often lacks descriptions at the shot level, making retrieval of relevant fragments at the shot level a time-consuming effort. Even if relevant video productions have been found, they have to be watched completely in order to narrow down the search selection to the relevant shots.

A promising approach to make search in video archives more efficient and effective is to develop automatic indexing techniques that produce descriptions at a higher semantic level that is better attuned to matching information needs. Such indexing techniques produce descriptions using a fixed vocabulary of so-called *high-level features* also referred to as *semantic concepts*. Typical examples of high-level features are objects such as 'car', persons such as 'Madeline Albright', scenes such as 'sky' or events like 'airplane takeoff'. These descriptors are named high-level features to make a clear distinction with low-level features such as colour, texture and shape. Low-level features are used as inputs for the detection of high-level features. In turn (and this is the main reason why they are called features), the high-level features can be used as features by a higher level interpretation module, combining different high-level features in a compositional fashion, e.g. 'car AND fire'.

Semantic concept indexing has been one of the objects of study of the TRECVid benchmarking evaluation campaign. More background about TRECVid is presented in Sections 2 and 3 of this chapter. Section 4 subsequently discusses the principal results and trends in the five iterations of the high-level feature detection task organized in each year during the period 2002-2006.

High-level feature detectors are usually built by training a classifier (often a support vector machine) on labeled training data. However, developing detectors with a high accuracy is challenging, since the number of positive training examples is usually rather small, so the classifier has to deal with class imbalance. There is also a large variation in example frames and the human labeling contains errors. From a development point of view, it is a challenge to find scalable methods that exploit multiple layers of rich representations and to develop fusion configurations that are automatically optimized for individual concepts. If the accuracy of such a detector is sufficiently high, it can be of tremendous help for a search task, especially if relevant concepts exist for the particular search query. For example, the performance of the query "Find two visible tennis players" benefits from using the high-level feature "tennis game". Of course the size of the concept lexicon and the granularity of the ontology it represents are seminal for the applicability of concept indexing for search. Over the last few years, the lexicon size of state-of-the-art systems for content based video access has grown from several tens to several hundreds and there is evidence that high-level features indeed improve search effectiveness and thus help to bridge the semantic gap.

However, there are several open research problems linked to using automatic semantic concept annotation for video search. Experience from five years of benchmarking high-level feature detectors at TRECVid has raised several issues. We mention a few here:

- The choice of a proper lexicon depends on the video collection and the envisaged queries, and no automatic strategy exists to assist in constructing such a lexicon.
- The accuracy of a substantial number of concepts is too poor to be helpful.

- The stability of the accuracy of concept detectors when moving from one collection to another has not been established yet.

Section 5 will discuss these and other open issues in some more detail and formulate an outlook on how to benchmark concept indexing techniques in the coming years.

2 Benchmarking Evaluation Campaigns, TREC, and TRECVid

The Text Retrieval Conference (TREC) initiative began in 1991 as a reaction to small collection sizes used in experimental information retrieval (IR) at that time, and the need for a more co-ordinated evaluation among researchers. TREC is run by the National Institute of Standards and Technology (NIST). It set out initially to benchmark the ad hoc search and retrieval operation on text documents and over the intervening decade and a half spawned over a dozen IR-related tasks including cross-language IR, filtering, IR from web data, interactive IR, high accuracy IR, IR from blog data, novelty detection in IR, IR from video data, IR from enterprise data, IR from genomic data, from legal data, from spam data, question-answering and others. 2007 was the 16th TREC evaluation and over a hundred research groups participated. One of the evaluation campaigns which started as a track within TREC but spawned off as an independent activity after 2 years is the video data track, known as TRECVid, and the subject of this paper.

The operation of TREC and all its tracks was established from the start and has followed the same formula which is basically:

- Acquire data and distribute it to participants;
- Formulate a set of search topics and release these to participants simultaneously and en bloc;
- Allow up to 4 weeks of query processing by participants and accept submissions of the top-1000 ranked documents per search topic, from each participant;
- Pool submissions to eliminate duplicates and use manual assessors to make binary relevance judgments;
- Calculate Precision, Recall and other derived measures for submitted runs and distribute results;
- Host workshop to compare results;

The approach in TREC has always been metrics-based — focusing on evaluation of search performance — with measurement typically being some variants of Precision and Recall.

Following the success of TREC and its many tracks, many similar evaluation campaigns have been launched in the information retrieval domain. In particular, in the video/image area there are evaluation campaigns for basic video/image analysis as well as for retrieval. In all cases these are not competitions with “winners” and “losers” but they are more correctly titled “evaluation campaigns” where interested parties can benchmark their techniques against others and normally they culminate in a workshop where results are presented and discussed. TRECVid is one such evaluation campaign and we shall see details of that in section 3, but first we shall look briefly at evaluations related to video processing.

ETISEO (Evaluation du Traitement et de l’Interprétation de Séquences Vidéo) [3] was an evaluation campaign that ran in 2005 and 2006. The aim was to evaluate vision techniques for video surveillance applications and it focussed on the treatment and interpretation of videos involving pedestrians and (or) vehicles, indoors or outdoors, obtained from fixed cameras. The video data used was single and multi-view surveillance of areas like airports, car parks, corridors and subways. The ground truth consisted of manual annotations and classifications of persons, vehicles and groups, and the tasks were detection, localization, classification and tracking of physical objects, and event recognition.

The PETS campaign (Performance Evaluation of Tracking & Surveillance) [6] is in its 10th year in 2007 and is funded by the European Union through the FP6 project ISCAPS (Integrated Surveillance of Crowded Areas for Public Security). PETS evaluates object detection and tracking for video surveillance, and its evaluation is also metrics based. Data in PETS is multi-view/multi-camera surveillance video using up to 4 cameras and the task is event detection for events such as luggage being left in public places.

The AMI (Augmented Multi-Party Interaction) project [2], funded by the European Union, targets computer enhanced multi-modal interaction, including the analysis of video recordings taken from multiple cameras, in the context of meetings. The project coordinates an evaluation campaign where tasks include 2D multi-person tracking, head tracking, head pose estimation and an estimation of the focus-of-attention (FoA) in meetings as being either a table, documents, a screen, or other people in the meeting. This is based on video analysis of people in the meeting and what is the focus of their gaze.

ARGOS [9] is another evaluation campaign for video content analysis tools. The set of tasks under evaluation have a lot of overlap with the TRECVid tasks and include shot bound detection, camera motion detection, person identification, video OCR and story boundary detection. The corpus of video used by ARGOS includes broadcast TV news, scientific documentaries and surveillance video.

Although even these evaluation campaigns in the video domain span multiple domains and genres as well as multiple applications, some of which are information retrieval, they have several things in common, including the following:

- they are all very metrics-based with agreed evaluation procedures and data formats;
- they are all primarily system evaluations rather than user evaluations;
- they are all open in terms of participation and make their results, and some also their data, available to others;
- they are all have manual self-annotation of ground truth or centralized assessment of pooled results;
- they all coordinate large volunteer efforts, many with little sponsorship funding;
- they all have growing participation;
- they all have contributed to raising the profile of their application and of evaluation campaigns in general;

What we can conclude from the level of activity in evaluation campaigns such as the above, and the TRECVid campaign which we will cover in the next section, is that they are established within their research communities as the means to carry out comparative evaluations.

3 The TRECVid Benchmarking Evaluation Campaign

The TREC Video Retrieval Evaluations began on a small scale in 2001 as one of the many variations on standard text IR evaluations hatched within the larger TREC effort. The motivation was an interest in expanding the notion of “information” in IR beyond text and the observation that it was difficult to compare research results in video retrieval because there was no common basis (data, tasks, or measures) for scientific comparison. TRECVid’s two goals reflected the relatively young nature of the field at the time it started, namely promotion of research and progress in video retrieval and in how to usefully benchmark performance. In both areas TRECVid has often opted for freedom for participants in the search for effective approaches over control aimed at finality of results. This is believed appropriate given the difficulty of the research problems addressed and the current maturity of systems.

TRECVid can be compared with more constrained evaluations using larger-scale testing as in the Face Recognition Grand Challenge (FRGC) [1] and in the context of benchmarking evaluation campaigns it is interesting to compare those in IR and image/video processing mentioned above, with such a “grand challenge”. The FRGC is built on the conclusion that there exist “three main contenders for improvements in face recognition” and on the definition of 5 specific conjectures to be tested. FRGC shares with TRECVid an emphasis on large data sets, shared tasks (experiments) so results are comparable, and shared input/output formats. But FRGC differs from TRECVid in that FRGC works with much more data and tests (complete ground truth is given by process of capturing data), more controlled data, focus on a single task, only non-interactive systems, and evaluation only in terms of verification and false accept rates. This makes it quite different from TRECVid.

The annual TRECVid cycle begins more than a year before the target workshop as NIST works with the sponsors to secure the video to be used and outlines associated tasks and measures. These are presented for discussion at the November workshop a year before they are to be used. They need to reflect interests of the sponsors as well as enough researchers to attract a critical mass of participants. With input from participants and sponsors, a set of guidelines is created and a call for participation is sent out by early February. The various sorts of data required are prepared for distribution in the spring and early summer. Researchers develop their systems, run them on the test data, and submit the output for manual and automatic evaluation at NIST starting in August. Results of the evaluations are returned to the participants in September and October. Participants then write up their work and discuss it at the workshop in mid-November – what worked, what didn’t work, and why. The emphasis in this is on learning by exploring. Final analysis and description of the work is completed in the months following the workshop and often include results of new or corrected experiments and discussion at the workshop. Many of the workshop papers are starting points for peer-reviewed publications, with a noticable effect on the scientific programme of multimedia conferences. Over the last few years, about 50 publications per year were reporting the use of a TRECVid test collection.

The TRECVid tasks which have been evaluated are shot boundary detection, detection of concepts or high-level features within shots, automatic detection of story bounds in broadcast TV news, three kinds of search (automatic, manual and interactive) and automatic video summarisation. In this chapter we gather together the work done and the contributions of the TRECVid high-level feature detection task since it started in 2002. We analyse its impact and we enlist what we believe to be the outstanding challenges and likely developments.

4 The TRECVid high-level Feature Detection Task

In this section we present a year-on-year summary of each of the first 5 iterations of TRECVid in terms of the feature detection task, tracking the development of the task and of system approaches over the years. Before we do that, however, we describe the overall task definition and the evaluation measures we used.

High-level feature extraction, important in its own right and a promising basis for search, was added at the urging of participants in 2002. The features tested have ranged over objects, people, and events with varying degrees of complexity that make some features very similar to topic text descriptions. Features have mostly been chosen with an eye to their likely use in search. Unlike search topics, feature definitions are known in advance of testing and contain only a short text description.

The TRECVid standard for correctness in annotation of feature training data and judging of system output is that of a human – so that examples which are very difficult for systems due to small size, occlusion, etc., are included in the training data and systems that can detect these examples get credit for them – as should be the case in a real system. This differs from some evaluations (e.g. FRGC) in which only a subset of examples that meet specified criteria are considered in the test. We want the TRECVid test collections to be useful long after the workshop and year in which they are created and even if systems improve dramatically.

Since in video there is no visual correlate of the word as an easily recognizable, reusable semantic feature, one of the primary hypotheses being examined in TRECVid is the idea that, given enough reusable feature detectors, such features might play something like the role words do in text IR. Of course, many additional problems - such as how to decide (automatically) which features to use in executing a given query – remain to be solved [7].

The task definition for high-level feature detection is as follows: given a collection of video, a set of high-level features and a common set of shot boundaries, return for each feature the list of the top video shots from the collection, ranked according to the highest probability of the shot containing the feature. In the initial year of TRECVid, the list was set to 1,000 though we later set the size of the submitted list to 2,000. Each participating group was allowed to submit multiple "runs", where each run was a variant of the technique they had developed for feature extraction.

One of the basic goals of TRECVid is to allow each group to compare the effectiveness of its algorithms on a given test set. This goal is more fundamental than the ability to compare two groups' systems or to measure the improvement of systems over time – even though the latter two capabilities are highly desirable.

A feature is defined as being present in the shot if it is present for at least 1 frame, so there is no concept of a feature being present for only part of a shot and the presence of each feature is assumed to be binary. Each of the submitted runs for each of the features from each of the participating groups were then pooled, duplicates removed, and the pool of shots was manually assessed for the presence/absence of the feature. Assuming the presence of a feature in a shot to be binary is a simplification adopted because of the benefits it afforded in pooling the results and in approximating recall and this was continued in all iterations of TRECVid.

When assessing the results of feature detection we employ the widely used *trec_eval* software to calculate standard information retrieval measures. Because in the initial years of the evaluation not all groups submitted results for all features, we calculated and presented the average precision for each feature for each group, rather than aggregating performance measures at the group or run levels. That changed in 2006.

One interesting aspect of the feature detection task was the provision of development data which could be used by participating groups to train their feature detection systems. The goal was to have groups train their systems on a common training dataset as that would afford more direct comparisons of systems and approaches rather than comparisons of the quality and amount of training data used. To that end we defined and provided sets of training data and encouraged participating groups to use this data for training their systems. In the latter years this training data consisted of manually annotated shots provided as part of large-scale community-based video annotation activities, an aspect of TRECVid which really allowed the benchmark to focus on system approaches rather than data availability.

4.1 TRECVid 2002

TRECVid 2002 was the second year of the TRECVid activity and at that time the evaluation was still one of the "tracks" of TREC [19]. One of the lessons we learned from the first year, in 2001, was that there is more to searching than the search task and that the automatic detection of some kind of semantic features was a necessary pre-requisite to any kind of quality retrieval. With that in mind we decided to include a task on automatic feature detection for shots and to schedule the submission deadline for this in such a way that the outputs of the feature detection task could be used by participants in the search task. At minimum, this would allow a group which was taking part in both feature detection and search to use the output of their own feature detectors in running their own search task. However, such is the spirit of co-operation in TRECVid that it became the norm that everybody's feature detection submissions are made available to every group taking part in the search task. This provides a great boost for the smaller groups or groups who wish to concentrate on the search task rather than on feature detection.

In 2002 we defined 10 features to be used, shown in column 1 of Table 4.1 with the feature number in parentheses, and these were suggested in on-line discussions among TRECVid participants. The video data in that year was taken from the Internet Archive and Open Video projects and slightly over 5 hours ($1,848$ shots in total) were used as test data for the feature detection task. The nature of the video was documentary and educational broadcast video. Common shot boundaries were provided by one of the participants and 9 groups of the 17 total participants in TRECVid 2002 submitted a total of 15 runs for assessment, and each was fully assessed manually for the presence of each of the 10 features.

Figure 1 presents the performance of runs which are at the median or above, submitted by participating groups, for each feature. Included as a dotted line in this figure is the baseline for a random feature detector. There is an artificial upper limit on average precision for some features (8 and 9) caused by the fact that groups submitted a maximum of 1,000 shots and these features have more ($1,382$ and $1,221$ respectively) features than could be identified in a single 1,000-shot submission. From this graph we can see that the performance on some features including *speech (8)*, *instrumental sound (9)* and *outdoors (1)*, is OK, while the absolute performance levels of others is poor, and none are great.

Summarising the approaches taken in 2002, we can say that they broadly fall into two approaches. The first of these was to hand-label shots for the occurrence of each of the features, to automatically extract low-level characteristics like colour, texture, motion, etc. from these shots, and then to train some automatic classifier like a support vector machine to recognise the high-level features from the low-level features. The second approach taken in 2002 was to exploit feature-specific characteristics and to use a face-recogniser to recognise faces, to use audio characteristics to recognise speech occurrence, etc. Of the two approaches, the first is the one which is scalable to large numbers of features and the one which has grown in importance since 2002. One other characteristic of the early feature detection approaches is that all approaches treated the detection of each feature as independent, something which has changed recently.

2002	2003	2004	2005	2006
Outdoors (1)	—	—	—	Outdoors (8)
Indoors (2)	Indoors (11)	—	—	—
Face (3)	—	—	—	Face (19)
People (4)	People (13)	People walk/run (35)	People walk/run (38)	Person (20) Walking/Running (24) People-Marching (35) Crowd (18) Government-Leader (21) Corporate-Leader (22) Police/Security (23) Military personnel (24)
Cityscape (5)	Building (14)	—	Building (42)	Building (9) Urban (16)
Landscape (6)	—	—	Mountain (44)	Mountain (12)
Text Overlay (7)	—	—	—	—
Speech (8)	Female Speech (18)	—	—	—
Instrumental Sound (9)	—	—	—	—
Monologue (10)	News subject monologue (21)	—	—	—
—	News subject face (12)	—	—	—
—	Road (15)	Road (37)	—	Road (13)
—	Vegetation (16)	—	—	Vegetation (11)
—	Animal (17)	—	—	Animal (26)
—	Car/truck/bus (19)	Train (31)	Car (47)	Car (30) Bus (31) Truck (32)
—	Aircraft (20)	Airplane takeoff (34)	—	Airplane (29)
—	Non-studio (22)	—	—	Studio (7)
—	Sports (23)	Basket score (33)	Sports (46)	Sports (1)
—	Weather news (24)	—	—	Weather news (3)
—	Zoom in (25)	—	—	—
—	Physical violence (26)	Physical violence (36)	—	—
—	Madeleine Albright (27)	Madeleine Albright (29)	—	—
—	—	Boats/ships (28)	—	Boat/Ship (33)
—	—	Bill Clinton (30)	—	—
—	—	Beach (32)	Waterscape/Waterfront (43)	Waterscape/Waterfront (17)
—	—	—	Explosion/Fire (39)	Explosion/Fire (36)
—	—	—	Map (40)	Maps (38) Charts (39)
—	—	—	U.S. Flag (41)	US flag (28)
—	—	—	Prisoner (45)	Prisoner (25)
—	—	—	—	Sky (14)
—	—	—	—	Snow (15)
—	—	—	—	Entertainment (2)
—	—	—	—	Desert (10)
—	—	—	—	Computer/TV-screen (27)
—	—	—	—	Natural-Disaster (37)

Table 1. Features Used in Various TRECVid campaigns. (Note that features were re-numbered starting at 1 in the 2006 TRECVid campaign.)

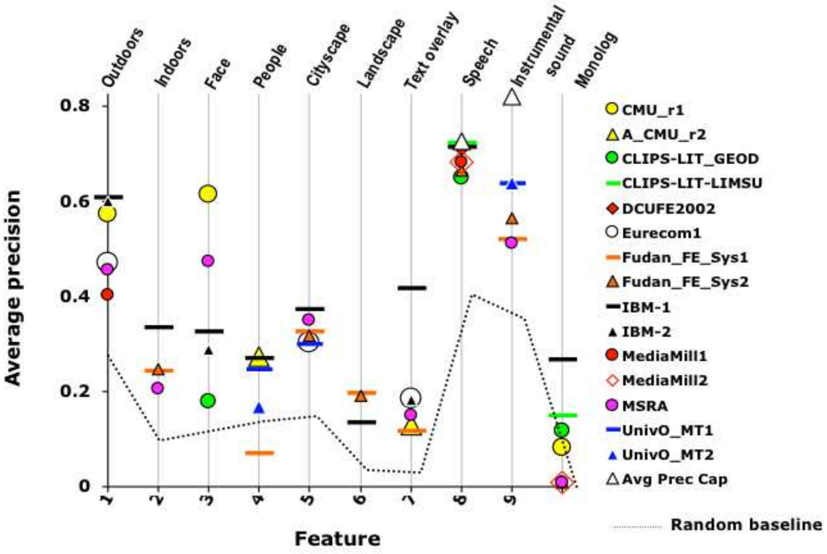

Fig. 1. Performance of high-level Feature Detection In TRECVid 2002

4.2 TRECVid 2003

In 2003, TRECVid branched out on its own and was not a track within the TREC framework, reflecting the importance that the activity was already having on the research field [18]. The track defined a set of 17 features to be detected, shown in the second column of Table 4.1, again with the feature numbers in parentheses. Some of these features were repeated from the previous year e.g. *indoors (11)* and *people (13)*, some were close variants such as *cityscape (5)* moving to *building (14)* and *speech (8)* moving to *female speech (18)*, and the rest were completely new. Of these new ones, some were predicted as being very difficult, such as *physical violence (26)* while others we could expect to be easier, such as *weather news (24)*. As in the previous year, evaluation was based on the average precision of each feature in each submitted run, and participants were asked to submit a list of up to 2,000 shots containing each of the features. The rest of the task was the same as in 2002.

The data used for feature detection in 2003 consisted of broadcast TV news from ABC, CNN and C-Span, and contained many advertisements and fast-moving shots associated with the news, as well much studio footage of anchorpersons. A total of 60 hours (32,318 shots) were used for the evaluation, a big step-up in size, and 10 groups submitted a total of 60 runs which were pooled and only partially assessed because of the large ramp-up in submissions and data volume from the data used in 2002.

Figure 2 shows the performance of the top 10 runs submitted for each of the 17 features. What is important in Figure 2 is not which group or which run led to the best performance, but the values of those best-performing runs for each feature. We can see that for some features like *weather news (24)* and *sporting event (23)* the best performance was excellent, that for many of the features the performance was mediocre, and that for some such as *non-studio*

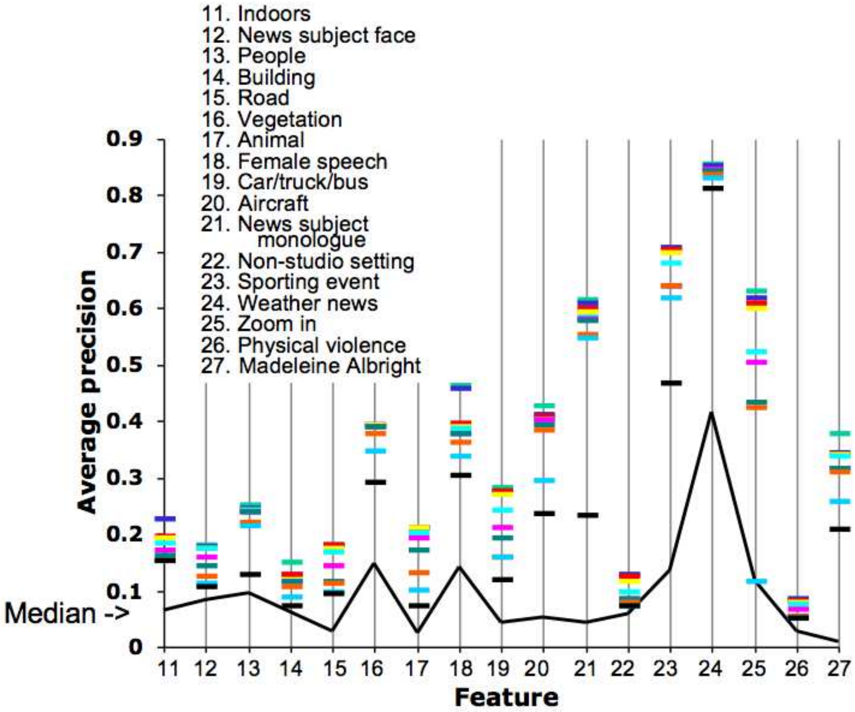

Fig. 2. Performance of Top-10 high-level Feature Detections per Feature in TRECVid 2003

setting (22) and *physical violence (26)* performance was quite poor. Figure 2 also shows the median performance across all 60 submitted runs.

One of the notable things about the 2003 evaluation was the introduction of a collaborative annotation of the training data allowing most of the participating groups to use the same training data. This activity was led by IBM Research and involved personnel from 23 groups volunteering to manually annotate some portion of the 60 hours of training data with 133 labels or semantic features [11]. Having this shared common training data, now with common annotations, allowed more groups to develop generic feature detection methods, for example using automatic machine learning techniques such as support vector machines. Also provided to the TRECVid organisers, and the participants, was the output of an automatic speech recognition system applied to both the training and test data, a facility which many groups used in developing their systems.

The main lesson learned from 2003 was that we could comfortably ramp up to 60 hours of test data and that there were enough groups able to complete the task on that volume of data to make it worthwhile. Also learned was the value of having a shared training dataset, in our case a shared annotation of video using semantic features.

4.3 TRECVid 2004

In 2004 the TRECVid feature detection task [10] used the same data source as in 2003 in order to minimise start-up costs for participants and to allow some stability for participants in completing the task. The track defined defined a set of 10 features to be detected, shown in the third column of Table 4.1 with the feature numbers in parentheses. Some of these features were repeated, e.g. *Road (37)*, some were variations such as *Airplane takeoff (34)* instead of *Aircraft (20)*, and others were new e.g. *Boats/ships (28)* and *Bill Clinton (30)*. Evaluation was the same as in 2003, based on average precision per feature calculated from submissions of up to 2,000 shots using a common shot boundary. Speech transcripts were also provided for the collection of 60 hours (33,367 shots) of broadcast TV news from the same TV sources as in 2003 used in the evaluation of feature detection. Training data provided to participants, who were encouraged to use this common training data, comprised the development data used in 2003, which had been annotated manually as part of a common annotation activity, and the submitted and evaluated runs from 2003.

Pooling of submitted runs was carried out differently to previous years. Each submitted run was divided into strata of depth 25 and the depth of the pools which were assessed manually, for each feature, varied according to the number of true shots found for that feature. This allowed us to spread the scarce resource of assessor time more evenly among the features to be assessed, in accordance with the "rate" at which shots truly containing each feature were found.

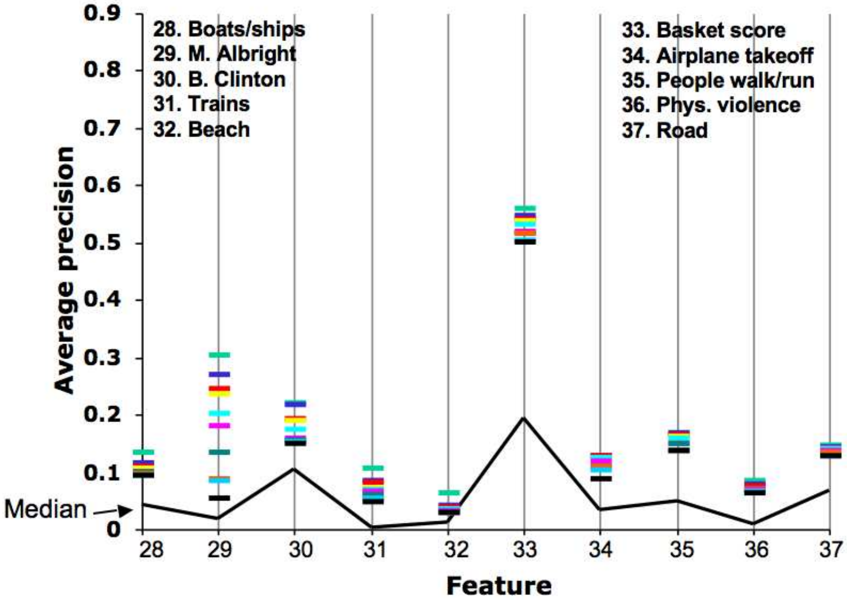

Fig. 3. Performance of Top-10 high-level Feature Detections per Feature in TRECVid 2004

A total of 83 runs were submitted by 12 groups who completed the feature detection task in 2004 and Figure 3 shows the performance of the top 10 runs submitted from among those 83 runs, for each feature across all submitted runs. Figure 3 also shows the median performance for each feature. Results across different features once again varied greatly in their mean and top performances. *Madeleine Albright (29)* in particular shows wide variability in detection performance while *Basketball score (33)* stands out with high scores. The latter feature was designed to be difficult (i.e. involving a scoring event), but turned out to be easy, since most basketball shots shown in news shows involve a score, so detection of basketball shots, for which discriminative textual features can be used, is sufficient.

Having some features common across the years allows direct comparison of approaches over time and we can see that the best run for *Madeleine Albright (27/29)* improved, while *Road (15/37)* and *Physical violence (26/36)* were about the same. For the rest of the features used in 2004 taken as a whole they were more difficult than those used in 2003. The features occurred also more rarely in the collection, the average rate of true shots per 1,000 test shots was 13 in 2004 versus 21 in 2003. However, an analysis of average precision versus number of true shots did not show a correlation.

Half of the 12 participating groups completed submissions for all 10 features, showing an increasing trend towards generic or feature-neutral approaches, which did have strong performance in terms of effectiveness. Most of the approaches taken by the groups addressed ways of combining or fusing the output of low-level feature analysis on the shots. Some used support vector machines, others used other approaches, such as Gaussian mixture models, etc.

4.4 TRECVid 2005

The search task in TRECVid in the years prior to 2005 had allowed a strong dependence on the automatic speech recognition provided to participants to evolve. In an attempt to move participating groups away from this dependency and to force greater concentration on exploiting the visual and audio aspects of video for video retrieval, TRECVid 2005 [16] introduced a new source of video data where the text was errorful and unreliable. Once again, as in 2003 and 2004, we used broadcast TV news but this time it was taken from English- Arabic- and Chinese-language sources, broadcast in November 2004. For each language we acquired automatic speech recognition output and for the non-English sources we provided machine translation into English. We again provided common shot boundaries, and common keyframes that groups could work with. The data used in the feature detection task consisted of 40 hours (45,765 shots), indicating a much higher shot change rate in the video than previously. Even though the genre of video was the same as previously, broadcast TV news, groups felt a need for a new source of training data for feature detection and another collaborative annotation of 39 features was completed [24], with 24 groups participating in the annotation activity using annotation tools donated by IBM Research and by Carnegie Mellon University. Almost all submitted runs (nearly 80%) used this training data alone.

The feature detection task used 10 features, selected by the organisers from an early version of the LSCOM (Large-Scale Concept Ontology for Multimedia) ontology which had been under construction specifically for broadcast TV news [13]. These are shown in the fourth column of Table 4.1 and included repeated, new and variation features. Evaluation metrics were the same as previously, though for this year all submissions for each feature were pooled and evaluated to a depth of 250 shots, requiring a total of 76,116 individual shot judgments.

Twenty-two groups completed the feature detection task in 2005 (a near doubling over the previous year) and they submitted a total of 110 runs, almost all of which contained submissions for each of the 10 features and Figure 4 shows the performance of the top 10 runs submitted for each feature. From that graph we can see that the scores for features are higher than in 2004, despite the fact that there are new data sources, and errorful text from speech

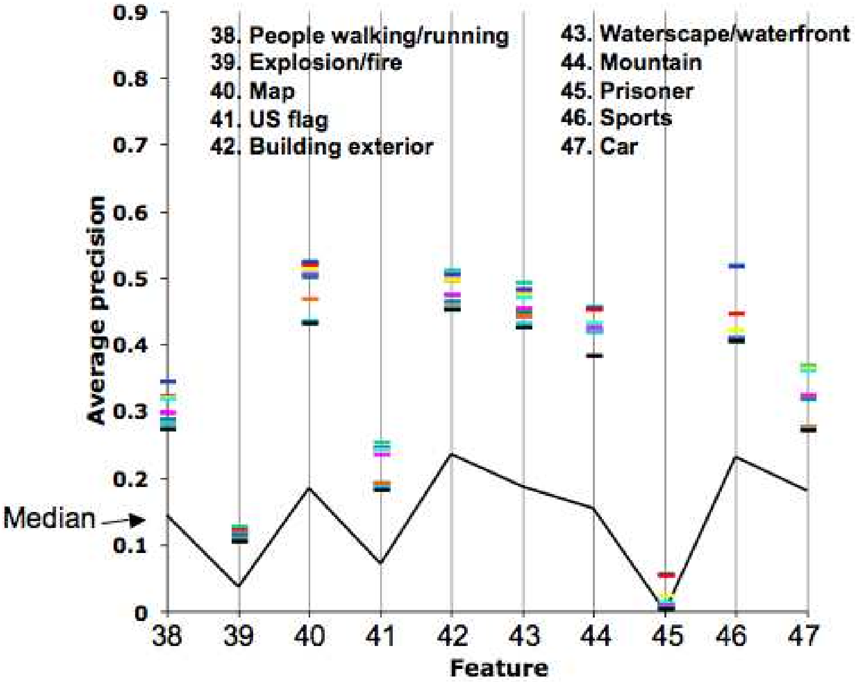

Fig. 4. Performance of Top-10 high-level Feature Detections per Feature in TRECVid 2005

recognition and machine translation. The sole feature that was common to both 2004 and 2005 (*People walking/running (35/38))* showed the same level of performance for the top-performing runs across the two years, though we cannot read too much into that because of the different data sources. The techniques used by participating groups continued to show the emergence of fusion methods, and some addressed the selection and combination of specific low-level visual features (colour, texture, etc.) for different semantic features.

4.5 TRECVid 2006

The fifth and final year of this retrospective on TRECVid feature detection [17] used the same data sources as in the previous year, broadcast TV news in three languages namely English (NBC, CNN and MSNBC), Arabic (LBC and HURRA) and Chinese (CCTV4. PHOENIX and NTDTV), taken from the same period in time. However one difference from the previous year was that the feature detection task was run on almost 150 hours of video or 79,484 individual shots. This represented a large step up in size.

In TRECVid 2006, participants were required to submit results for all 39 individual features defined by the LSCOM workshop as “LSCOM-Lite”, rather than some self-selected subset thereof. This was intended to further promote the use of generic means of training feature detectors. Of these 39 features, the number to be evaluated manually was at first kept small (10) so as to make the assessment manageable in this iteration of TRECVid. However, work by Yilmaz and Aslam [25] had resulted in methods for estimating standard system performance measures using relatively small samples of the usual judgment sets so that larger numbers of features could be evaluated using the same amount of judging effort. Using the TRECVid 2005 high-level feature task results, an analysis of the new estimate for average precision — inferred average precision (infAP) — at various levels of judgment sampling (80%, 60%, 40% and 20%) showed very good estimation of average precision in terms of actual values of the measures. Thus we decided to use this approach in evaluating feature detection performance in TRECVid 2006 using a 50% sample of the usual feature task judgment set and this allowed us to judge 20 features from each group rather than the 10 initially planned. For each feature, all submissions down to a depth of at least 100, and 145 on average, were added to the pools from which a 50% sample was judged.

Participants in the evaluation were given access to the MediaMill challenge data which included 101 low-level features as well as estimated 101 MediaMill high-level concept occurrences in the 2005 and 2006 test data, plus the manual annotations on the development data for 449 features taken from the fuller LSCOM ontology which had also been provided in the previous year as training data. That meant that there was a total of 159 hours of annotated training/development data for those who needed it ! As usual we also provided common shot boundaries and common keyframes for groups to work on. Throughout the previous iterations of the feature detection task most groups had come to depend on the keyframe as the shot representative and had applied their feature detection techniques to the keyframe rather than the whole shot. As we know, the definition of a feature occurrence in a shot is that the feature can occur *anywhere* in the shot, and participants take something of a risk by working on the keyframe only. In 2006, however, 8 of the 30 teams completing the task looked at more than just the keyframe in deciding whether a feature was present in the shot, and this is a useful development which we expect to see continued.

The TRECVid feature detection task in 2006 resulted in 30 groups submitting a total of 125 runs and the top 10 runs for each of the 20 features that were evaluated are shown in Figure 5. Of these submitted runs, support vector machines are still the dominant classifier with robust results. Good systems were those which combined representations at multiple granularities (local, regional, global) with use of salient point representations gaining ground. Good systems also combined various types of features (colour, texture, shape, edges, acoustic, face, text). Many interesting multi-modal and concept fusion experiments were carried out and multi-concept fusion still seems of limited use, perhaps because there are not enough concepts that support each other in the relatively small set of 39 used in 2006.

4.6 Data

Data is the element of the evaluation with the fewest degrees of freedom. While one can ruminate about ideal test collections, in practice when organising resources for a large evaluation like TRECVid one more often takes what one can get – if it can at all be useful – and acquisition of video data from content providers has always been difficult in TRECVid. TRECVid has formally evaluated systems against produced video and in 2005 and 2006 has explored tasks against unproduced, raw video as well.

Produced video

From the 11 hours of video about NIST used for a feasibility test in 2001, TRECVid moved in 2002 to 73 hours of vintage video mainly from the Internet Archive [4] – a real collection still needing a search engine to find video for re-use. Participants downloaded the data themselves.

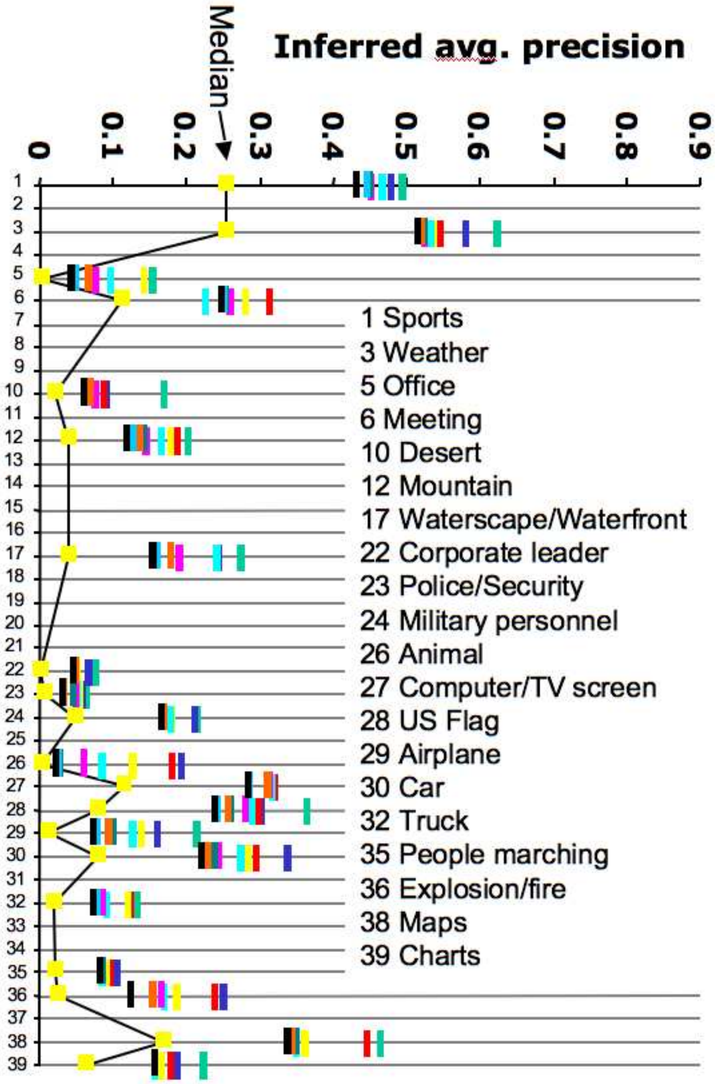

Fig. 5. Performance of Top-10 high-level Feature Detections per Evaluated Feature in TRECVid 2006

Then in 2003 TRECVid began working on broadcast TV news video from a narrow time interval – a new genre, much more consistent in its production values than the earlier data, and larger in size. Data set sizes made it necessary to ship the video on hard drives – a method that has worked well with the exception of one year in which groups with older versions of the Windows operating system could could not access drives of the size used.

Another important change was the shift to two-year cycles. Within the same genre enough data was secured so that training and test data could be provided in the first year, with the training data annotated and re-used in the second year during which only new test data would be provided. This reduced the overhead of system builders adapting to new video, reduced the overhead of training data annotation and maximized its use, and removed a “new genre” factor from influencing results in the second year of the cycle. TRECVid 2006 completed the second such two-year cycle. Test/training data amounts (in hours) have grown as follows: 2003 (66/67), 2004 (70/0), 2005 (85/85), 2006 (158/0). The video in 2003-2004 was from English-speaking sources. In 2005 and 2006 Chinese- and Arabic-speaking sources were added to the mix. Automatic machine translation was used to get English text from Chinese and Arabic speech.

We have learned that broadcast news video has special characteristics with consequences for the evaluation and systems used to detect features. It is highly produced, dominated by talking heads, and contains lots of duplicate or near duplicate material. Highly produced news video exhibits production conventions that systems will learn but with negative consequences when detectors learned on one news source are applied to another with different production conventions. This a real problem which systems need to confront and makes it important that the training data come from multiple sources. There are 8 different sources and 11 different programs in the 2006 test data and a significant number of test data sources did not occur in the training data.

Much of broadcast TV news footage is visually uninformative as the main information is contained in the reporter’s or anchorperson’s speech. This makes the TRECVid search task more difficult because the topics ask for *video* of objects, people, events, etc. *not information* about them. Video of a reporter talking about person X does not by itself satisfy a topic asking for video of person X. The search task is designed this way because it models one of two work situations. One is a searcher looking at open source video, interested in objects, people, events, etc. that are visible but not the subject the speech track, in the unintended visual information content about people, infrastructure, etc. The other situation is a video producer looking for clips to “re-purpose”. The original intent often reflected in the speech track is irrelevant. Of course, the speech track (or text from speech) can be very helpful in finding the right neighborhood for browsing and finding the video requested by some topics. But even when speech about X is accompanied by video of X they tend to be offset in time.

Highly produced news video also exhibits lots of duplicate or near duplicate segments due to repeated commercials, stock footage, previews of coming segments, standard intro and exit graphics, etc. Measuring the frequency of various sorts of duplicates or near duplicates is an unresolved research issue, as is assessing the distorting effect they may have on basic measures such as precision and recall.

4.7 Measurements

The TRECVid community has not spent significant amounts of time debating the pros and cons of various measures of effectiveness as applied to the feature detection task. Instead we have profited by battles fought long ago in the text IR community. While the choice of a single number (average precision or inferred average precision) to describe generalized system performance is as useful (e.g., for optimization, results graphs) as it is restrictive, TRECVid continues the TREC tradition of providing various additional views of system effectiveness for their diagnostic value and better fit for specific applications and analyses.

For feature extraction TRECVid adopted the family of precision- and recall-based measures for system effectiveness that have become standard within the TREC (text) retrieval community. Additional measures of user characteristics, behavior, and satisfaction developed by the TREC interactive search track over several years were adopted for use by interactive video search systems.

5 Discussion: achievements and future directions

The importance of feature detection as a task has grown over the years. The number of groups working on features benchmarking increased over the years from 9, 10, 12 and 22 to 30 in 2006. Now more groups participate in high-level feature task than in the search task. In the first few years, the potential use of concept detectors for search was only a hypothesis, and it was unclear whether a generic approach to building concept detectors would be feasible. A number of recent studies did specifically study this hypothesis. Indeed, experiments show that there is a clear positive correlation between search performance and the number of available high-level concept detectors[21]; experiments also suggest a benefit of combining multiple detectors[15, 21]. Concepts can mutually reinforce each other (i.e. a concept detector can fruitfully exploit the output of other detectors for related concepts). The targeted effort to design a concept ontology for broadcast news, LSCOM [5], has also been very influential, since it created the possibility to use the semantic relations between concepts for the search task. The impact of the use of concepts on search performance depends on the accuracy of their corresponding detectors. A recent simulation study shows that even the use of many low accuracy detectors can improve search results substantially, provided the lexicon consists of several thousand concepts [8]. It is important to realize though that designing and evaluating concept ontologies is still an open research problem. LSCOM has been designed for broadcast news. Other video genres such as home video, drama or surveillance video require alternative ontologies. An example is the recently announced ontology for home video [12], with a focus on family life and travel scenes.

Five years of feature detection benchmarking have resulted in a certain consolidation with regards to the main architecture/approach for building semantic concept detectors. Most TRECVid systems have from the beginning treated feature detection as a supervised pattern classification task based on one key frame for each shot. They have been converging on generic learning schemes over handcrafted detector construction as the desire to increase the set of features to a hundred and an order of magnitude beyond [5] has grown and scalability of learning scheme becomes critical. The TRECVid 2006 feature task recognized this by requiring submissions for 39 features of which 20 were evaluated.

Naphade and Smith [14] surveyed successful approaches for detection of semantic features used in TRECVid systems and abstracted a common processing pipeline including feature extraction, feature-based modeling (using e.g., Gaussian mixture models, support vector machines, hidden Markov models, and fuzzy K-nearest neighbors), feature-specific aggregation, cross-feature and cross-media aggregation, cross-concept aggregation, and rule-based filtering. This pipeline may accommodate automatic feature-specific variations [22]. They documented over two dozen different algorithms used in the various processing stages and note a correlation between number of positive training examples and best precision at 100.

Snoek et al. [23] identified a common architecture for TRECVid feature extraction systems and experimental variations including unimodal versus multimodal content analysis and early versus late fusion. They also point out some of the implications of the choices made and cite evidence that different features may require different approaches.

Beyond the above generalizations, conclusions about relative effectiveness of various combinations of techniques are generally possible only in the context of a particular group's experiments as described in their site reports in the TRECVid proceedings. In 2005, groups found evidence for the value of local over global fusion, multilingual over monolingual runs, multiple over single text sources (Carnegie Mellon University), parts-based object representation (Columbia University), various fusion techniques across features and learning approaches (IBM), automatically learned feature-specific combinations of content, style, and context analysis, and a larger (101) feature set (University of Amsterdam).

Still, each year the question is faced of how the community as a whole can learn from 100+ experiments. Tradition has always been to minimize control over the structure of the experiments for individual sites. Now approaches are consolidating, and it may become more attractive to control more factors in the experimental setting in order to make submissions more comparable across sites. One way to facilitate meta-analysis of experiment results across sites is to classify systems based on an ontology of experimental choices that has been constructed for the design of a concept detector architecture. A preliminary classification scheme was used for a post hoc classification of runs during TRECVid 2006 [17]. A more principled submission metadata annotation effort is underway for TRECVid 2007. A more rigorous way to focus (part of) the attention of the entire TRECVid on a particular research question is to include one or more required runs, that have to meet a particular condition e.g. use of a particular dataset for training, or are limited to visual only information. Another option would be to settle on an agreed (baseline) architecture and set of components in order to reduce the number of factors affecting results. So far, TRECVid has focused on facilitating a benchmark infrastructure for within-site experiments, and a gradual change towards support for across site comparisons to facilitate meta-studies of important well-defined research hypotheses is contemplated in order to enable the quantification of the relative utility of the many techniques that are studied in the high-level feature task.

The basics of the high-level feature benchmark have only changed gradually over the years. It is important to provide the community with a stable forum, which enables the measurements of year to year increments. Still, every year some small changes are introduced to increase the effectiveness and generalizability of the experiments, while maximizing the impact of the limited assessment time. The choice of features itself has always been difficult, since the test collection needs enough positive examples present, but not too many. The feature choice has always been balanced across several categories (person, object, event). Especially the definition of event features have been refined in order to stimulate work on the analysis of complete shots rather than keyframes. For TRECVid 2006, 8 groups did analyze more than just one keyframe per shot. In the broadcast news domain, shots are fairly short, for longer shots, it might make sense to annotate the presence of a feature at the frame level.

A recurring question is whether the performance of high-level feature detection is improving over the years. This cannot be answered, since the dataset changes every year, even if some features remain in the task. In order to make such a comparison we would at least need to have a reference system run on both test datasets to gauge the difference in the test data. In addition the use of average precision is sometimes questioned. The advantage of mean average precision is that it is a stable measure, which cannot be said about precision oriented measures such as precision@10. The high-level feature detection task in itself is an example of an intrinsic evaluation, where the performance is measured with respect to a ground truth. Features are important for search. But how can features effectively be used by users during a search task,and do we deal with varying accuracies of detectors ? There are many important aspects related to the actual use of concept detectors for search that are not measured by the current task.

Looking ahead, there are still many open issues and challenges before large scale concept detection is well understood. We mention a few:

- What are the limits on the generalizability of detectors, i.e., how reusable are the detectors, and how can we measure this in an affordable way given the further constraint that changing data sets is expensive ?
- What are the issues in moving from one video genre to another video genre — can it be done ?
- What are the necessary sizes for a training set for HLF machine learning ? What is the necessary accuracy for manual annotation of development data ?
- Are there issues of execution speed for HLF detection, especially if we want to scale up to very large-sized archives. Can we do this on-the-fly, at search time ?

In summary, the TRECVid high-level feature detection task has realized an important testbed for concept detection architectures that have proven to be an important performance enhancing component in video search systems. By maintaining a balance between stability and gradual refinements in the experimental set-up, this task has allowed the community to make increments in scale and (although harder to measure) detector effectiveness. Future experiments should be more focused on quantifying the robustness of the technology, how well can detectors be applied in different domains, and on better comparability of the experiments across sites and across collections in order to answer community-wide high-level research questions.

Acknowledgement. There are many people and groups of people whom the authors would like to thank for making the TRECVid activity possible. These include the funders, DTO, the organisers of the collaborative video annotation Cheng-Yun Li and Timo Volkmer at IBM Research and Alex Hauptmann of Carnegie Mellon University, shot bound detection donations from CLIPS and HHI, to KFs, from DCU, to BBC and Open Video for data donations, to LIMSI, CMU, BBN and LDC for ASR donations, to various groups for donating data and supplementary info incl MediaMill and LSCOM.

References

1. Face Recognition Grand Challenge. URL:www.frvt.org/FRGC, 2006.
2. AMI: Augmented Multi-Person Interaction. URL:www.amiproject.org/, Last checked 9 September 2007.
3. ETISEO: Video Understanding Evaluation. URL:www.silogic.fr/etiseo/, Last checked 9 September 2007.
4. The Internet Archive Movie Archive home page, Last checked 14 September 2007.
5. LSCOM Lexicon Definitions and Annotations. URL:www.ee.columbia.edu/dvmm/lscom, Last checked 14 September 2007.
6. PETS: Performance Evaluation of Tracking and Surveillance. URL:www.cvg.cs.rdg.ac.uk/slides/pets.html, Last checked 9 September 2007.
7. M. G. Christel and A. G. Hauptmann. The Use and Utility of High-Level Semantic Features in Video Retrieval. In *Proceedings of the International Conference on Video Retrieval*, pages 134–144, Singapore, 20-22 July 2005.
8. A. Hauptman. How many high-level concepts will fill the semantic gap in video retrieval? In *proceedings of the ACM International Conference on Image and Video Retrieval*, 2007.
9. P. Joly, J. Benois-Pineau, E. Kijak, and G. Quénot. The ARGOS campaign: Evaluation of video analysis and indexing tools. *Image Commun.*, 22(7-8):705–717, 2007.

10. W. Kraaij, A. F. Smeaton, P. Over, and J. Arlandis. TRECVID 2004—An Overview. In *Proceedings of the TRECVID Workshop (TRECVID 2004)*, Gaithersburg, MD, November 2004.
11. C.-Y. Lin, B. L. Tseng, and J. R. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. *Proceedings of the TRECVID 2003 Workshop*, 2003.
12. A. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa. Kodak's consumer video benchmark data set: concept definition and annotation. In *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 245–254, New York, NY, USA, 2007. ACM Press.
13. M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. G. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia Magazine*, 13(3):86–91, 2006.
14. M. R. Naphade and J. R. Smith. On the Detection of Semantic Concepts at TRECVID. In *MULTIMEDIA'04: Proceedings of the 12th ACM International Conference on Multimedia*, pages 660—667, New York, NY, USA, 10-16 October 2004.
15. A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 991–1000, New York, NY, USA, 2007. ACM Press.
16. P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton. TRECVID 2005—An Overview. In *Proceedings of the TRECVID Workshop (TRECVID 2005)*, Gaithersburg, MD, November 2005.
17. P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton. TRECVID 2006—An Overview. In *Proceedings of the TRECVID Workshop (TRECVID 2006)*, Gaithersburg, MD, November 2006.
18. A. F. Smeaton, W. Kraaij, and P. Over. TRECVid 2003: An overview. In *TREC2003: Proceedings of the TREC Workshop (TREC 2003)*, Gaithersburg, MD, November 2003.
19. A. F. Smeaton and P. Over. The TREC-2002 video track report. In *TREC2002: Proceedings of the TREC Workshop (TREC 2002)*, Gaithersburg, MD, November 2002.
20. A. 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 Recognition and Machine Intelligence*, 22(12):1349–1380, 2000.
21. C. G. Snoek and M. Worring. Are concept detector lexicons effective for video search?
22. C. G. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. Smeulders. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE Transactions, PAMI*, 28(10):1678–1689, 2006.
23. C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, Santa Barbara, CA, USA, 2006. ACM Press.
24. 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 *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 892–901, Hilton, Singapore, 2005. ACM Press.
25. E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 102–111, Arlington, Virginia, USA, 2006. ACM Press.