TRECVID 2005 - An Overview

Paul Over[◇], Tzveta Ianeva[◇†], Wessel Kraaij[‡], and Alan F. Smeaton[○]

[◇]Retrieval Group
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA
{over,tianeva}@nist.gov

[○]Adaptive Information Cluster /
Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland
asmeaton@computing.dcu.ie

[‡]TNO Information and Communication Technology
Delft, the Netherlands
wessel.kraaij@.tno.nl

[†]Departament d' Informàtica
Universitat de València
València, Spain
tzveta.ianeva@uv.es

March 27, 2006

1 Introduction

TRECVID 2005 represented the fifth running of a TREC-style video retrieval evaluation, the goal of which remained to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. Over time this effort is yielding a better understanding of how systems can effectively accomplish such retrieval and how one can reliably benchmark their performance. TRECVID is funded by the Disruptive Technology Office (DTO) and the National Institute of Standards and Technology (NIST).

Forty-two teams from various research organizations[1] — 11 from Asia/Australia, 17 from Europe, 13 from the Americas, and 1 US/EU team — participated in one or more of five tasks: shot boundary determination, low-level feature (camera motion) extraction, high-level feature extraction, search (automatic, manual, interactive) or pre-production video management. Results for the first four tasks were scored by NIST using manually created truth data for shot boundary determination and camera motion detection. Feature and search submissions were evaluated based on partial manual judgments of the pooled submissions. For the fifth exploratory task participants evaluated their own systems.

Test data for the search and feature tasks was about 85 hours of broadcast news video in MPEG-1 format from US, Chinese, and Arabic sources that had been collected in November 2004. Several hours of NASA's Connect and/or Destination Tomorrow series which had not yet been made public were provided by NASA and the Open Video Project for use along with some news video in the shot boundary task test collection. The BBC provided 50 hours of "rushes" - pre-production travel video material with natural sound, errors, etc. - against which participants could experiment and try to demonstrate functionality useful in managing and mining such material.

This paper is an introduction to, and an overview of, the evaluation framework — the tasks, data, and measures. The results, and the approaches taken by the participating groups. For detailed information about the approaches and results, the reader should see the online proceedings on the TRECVID website (www-nlpir.nist.gov/projects/trecvid)

1.1 New in TRECVID 2005

While TRECVID 2005 continued to work primarily with broadcast news, the addition of sources in

[1]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.

Arabic and Chinese complicated the already difficult search and feature detection tasks by introducing greater variety in production styles and more errorful text-from-speech due at least to the addition of fully automatic translation to English for the Arabic and Chinese sources.

A new low-level feature (camera motion) detection task was piloted in 2005. This task turned out to be quite problematic to run, as is explained in the section on that task but the quality of the results is impressive indicating that camera motion detection can be done accurately.

The BBC rushes presented special challenges (e.g., video material with mostly only natural sound, errors, lots of redundancy) and a special opportunity since such material is potentially valuable but currently inaccessible.

There was an increase in the number of participants who completed at least one task - up to 42 from last year's 33. See Table 1 for a list of participants and the tasks they undertook.

2 Data

2.1 Video

The total amount of news data for the evaluated tasks was about 169 hours of video: 43 in Arabic, 52 in Chinese, 74 in English. These data were collected by the Linguistic Data Consortium during November of 2004, digitized, and transcoded to MPEG-1.

A shot boundary test collection for 2005, comprising about 7 hours, was drawn at random from the total news collection. To these were added 4 NASA science videos. It then comprised 12 videos (8 news, 4 NASA) for a total size of about 4.64 gigabytes. The characteristics of this test collection are discussed below. The shot boundary determination test data were distributed by NIST on DVDs just prior to the test period start.

The total news collection minus the shot boundary test set was divided roughly in half chronologically. The earlier half was provided as development data for the high-level feature task as well as the search task. The later half was used as test data. Both the development and test data were distributed on hard disk drives by LDC.

Table 2: News files provided

Required	Development				Test			
	MPEG-1	Virage ASR/MT	MS-ASR	XLT of MS-ASR	MPEG-1	Virage ASR/MT	MS-ASR	XLT of MS-ASR
Ara	26	26	--	--	30	30	--	--
Chi	43	42	--	39	42	42	--	41
Eng	68	--	68	--	68	--	68	--

Optional	Development				Test			
	MPEG-1	Virage ASR/MT	MS-ASR	XLT of MS-ASR	MPEG-1	Virage ASR/MT	MS-ASR	XLT of MS-ASR
Chi	43		39		42		42	
Eng	68	57	40		68	39	42	

2.2 Common shot reference, keyframes, speech transcripts

The entire feature/search collection was automatically divided into shots by Christian Petersohn at the Fraunhofer (Heinrich Hertz) Institute in Berlin. These shots served as the predefined units of evaluation for the feature extraction and search tasks. The feature/search test collection contained 140 files/videos and 45,765 reference shots.

A team at Dublin City University's Centre for Digital Video Processing extracted a keyframe for each reference shot and these were made available to participating groups.

Carnegie Mellon University (CMU) provided the output of the beta version of a Microsoft Research automatic speech recognition system (ASR) for the English news sources, as well as ASR output for the Chinese files and machine translation (MT)(Vogel et al., 2003) of that output to English.

A contractor for the US Intelligence Community provided ASR/MT output for the Arabic files. They also produced ASR/MT for the Chinese files and this was made optionally available. While the ASR/MT provided by the contractor is the output of a commercial software on real data (Virage VideoLogger, Language Weaver), the system was not tuned to the TRECVID data and the contractor was not able to track down and fix errors that may have occurred in the processing.

See Table 2 for a summary of the files and file types provided.

2.3 Common feature annotation

In 2005 each of about 100 researchers from some two dozen participating groups annotated a subset of some 39 features in the development data using a tool developed by CMU or a new one from IBM. The total set of annotations was distributed to all groups that contributed – for use in training feature detectors and search systems.

In order to help isolate system development as a factor in system performance each feature extraction task submission, search task submission, or donation of extracted features declared its type:

A - system trained only on common TRECVID development collection data, the common annotation of such data, and any truth data created at NIST for earlier topics and test data, which is publicly available. For example, common annotation of 2003 training data and NIST's manually created truth data for 2003 and 2004 could in theory be used to train type A systems in 2005.

B - system trained only on common development collection but not on (just) common annotation of it

C - system is not of type A or B

Since by design there were multiple annotators for most of the common training data features but it was not at all clear how best to combine those sources of evidence, it seemed advisable to allow groups using the common annotation to choose a subset and still qualify as using type A training. This was the equivalent of adding new negative judgments. However, no new positive judgments could be added.

3 Shot boundary detection

Movies on film stock are composed of a series of still pictures (frames) which, when projected together rapidly, the human brain smears together so we get the illusion of motion or change. Digital video is also organized into frames - usually 25 or 30 per second. Above the frame, the next largest unit of video both syntactically and semantically is called the shot. A half hour of video, in a TV program for example, can contain several hundred shots. A shot was originally the film produced during a single run of a camera from the time it was turned on until it was turned off or a subsequence thereof as selected by a film editor. The new possibilities offered by digital video have blurred this definition somewhat, but shots, as perceived by a human, remain a basic unit of video, useful in a variety of ways.

The shot boundary task is included in TRECVID as an introductory problem, the output of which is needed for most higher-level tasks. Groups can work for their first time in TRECVID on this task, develop their infrastructure, and move on to more complicated tasks the next year, or they can take on the more complicated tasks in their first year, as some do. Information on the effectiveness of particular shot boundary detection systems is useful in selecting donated segmentations used for scoring other tasks.

The task was to find each shot boundary in the test collection and identify it as an abrupt or gradual transition, where any transition, which is not abrupt is considered gradual.

3.1 Data

The shot boundary test videos contained 744,604 total frames (20% more than last year) and 4,535 shot transitions (5.6% fewer than last year).

The reference data was created by a student at NIST whose task was to identify all transitions and assign each to one of the following categories:

cut - no transition, i.e., last frame of one shot followed immediately by the first frame of the next shot, with no fade or other combination;

dissolve - shot transition takes place as the first shot fades out *while* the second shot fades in

fadeout/in - shot transition takes place as the first shot fades out and *then* the second fades in

other - everything not in the previous categories e.g., diagonal wipes.

Software was developed and used to sanity check the manual results for consistency and some corrections were made. Borderline cases were discussed before the judgment was recorded.

The freely available software tool [2] VirtualDub was used to view the videos and frame numbers. The distribution of transition types was as follows:

- 2,759 — hard cuts (60.8%)
- 1,382 — dissolves (30.5%)
- 81 — fades to black and back (1.8%)

[2]The VirtualDub (Lee, 2001) website contains information about VirtualDub tool and the MPEG decoder it uses.

- 313 — other (6.9%)

3.2 Evaluation and measures

Participating groups in this task were allowed up to 10 submissions and these were compared automatically to the shot boundary reference data. Each group determined the different parameter settings for each run they submitted. Twenty-one groups submitted runs.

Detection performance for cuts and for gradual transitions was measured by precision and recall where the detection criteria required only a single frame overlap between the submitted transitions and the reference transition. This was to make the detection independent of the accuracy of the detected boundaries. For the purposes of detection, we considered a submitted abrupt transition to include the last pre-transition and first post-transition frames so that it has an effective length of two frames (rather than zero).

Analysis of performance individually for the many sorts of gradual transitions was left to the participants since the motivation for this varies greatly by application and system.

Gradual transitions could only match gradual transitions and cuts match only cuts, except in the case of very short gradual transitions (5 frames or less), which, whether in the reference set or in a submission, were treated as cuts. We also expanded each abrupt reference transition by 5 frames in each direction before matching against submitted transitions to accommodate differences in frame numbering by different decoders.

Accuracy for reference gradual transitions successfully detected was measured using the one-to-one matching list output by the detection evaluation. The accuracy measures were frame-based precision and recall. Note that a system could be very good in detection and have poor accuracy, or it might miss a lot of transitions but still be very accurate on the ones it finds.

3.3 Approaches in brief

The *City University of Hong Kong* used spatio-temporal (SD) slides, which are time vs. space representations of video. Shot transition types (cuts, dissolves) appear in SDs with certain characteristics. Gabor features were used for motion texture and SVMs for binary classification. They expanded on an existing approach by including flash detection and extra visual features to discriminate gradual transitions. Because of image processing and use of support vector machines (SVM), the approach is computationally expensive. The *CLIPS-IMAG, LSR-IMAG, NII* approach was essentially a rerun of their 2004 system, which may offer some insight into the relative difficulty of the 2005 test data compared to that from 2004. Cuts were detected by image comparisons after motion compensation and gradual transitions by comparing norms of first and second temporal derivatives of the images. Performance was about real-time, good on gradual transitions.

Fudan University approached the task using frame-frame similarities, varying thresholds, and SVM classifiers. They explored HSV (hue, saturation, value) vs. CIE L*a*b* color spaces. The Fudan system classified short gradual transitions as cuts. This differs from the TRECVID definition, depressing results. Performance was in the middle in runtime and in accuracy. The team at *FX Palo Alto* built on previous years with intermediate visual features derived from low-level image features for pairwise frame similarities over local and longer-distances. The system used the similarities as input to a k-nearest neighbor (kNN) classifier, and added information-theoretic secondary feature selection to select features used in classifier. Feature selection/reduction yielded improved performance but not as good as expected because of sensitivity to the training data.

Hong Kong Polytechnical University computed frame-frame similarities over different distances and generated distance maps, which have characteristics for cuts, gradual transitions, flashes, etc. Performance was about equal to real-time. The researchers at *IBM* built upon previous CueVideo work at TRECVID. The system was the same as 2005, except it used a different video decoder to overcome color errors. Switching the video decoder yielded improved performances. They noticed that the TRECVid 2005 video encoding had no B-frames. At *Imperial College London* the approach was the same as previous TRECVid submissions – exploiting; frame-frame differences based on color histogram comparisons.

The *Indian Institute of Technology's* system focused on hard cuts only. It addressed false positives caused by abnormal lighting (flashes, reflections, camera movements, explosions, fire, etc.) A 2-pass algorithm - first computed similarity between adjacent frames using wavelets, then focused on candi-

date areas to eliminate false positives. Computation time was about the same as real-time. The team at *KDDI* developed a system that worked in the compressed domain and so was fast. Luminance adaptive thresholds and image cropping yielded good results. They extended last year's work by adding edge features from discrete cosine transform (DCT) image, color layout, and SVM learning. *LaBRI* from the University of Bordeaux used last year's approach in the compressed domain, computed motion and frame statistics, then measured similarity between compensated adjacent I-frames. Performance was good on hard cuts, and fast; but not so on gradual transitions.

Two teams participated as category C teams, meaning that they are unable to provide details about their systems. The *Motorola Multimedia Research Laboratory* submitted a run. The system execution was fast. The *National ICT Australia* system used video analysis and machine learning. The computation involved was relatively expensive.

RMIT created a new implementation of their sliding query window approach, computing frame similarities among X frames before/after based on color histograms. They experimented with different (HSV) color histogram representations.; Feature selection/reduction yielded improved performances. Performance was not as good as expected because of sensitivity to the training data; The system developed at the *University of Delft* represented video as spatio-temporal video data blocks and extracted patterns from these to indicate cuts and gradual transitions. The approach was efficient and is likely to be expanded to include camera motion information.

At *Tsinghua University* researchers reimplemented previous years' very successful approaches, which had evolved to a set of collaboration rules for various detectors. The new system is a unified framework with SVMs combining fade-in/out detectors, gradual transition detectors and cut detectors, each developed in previous years; Despite individual detectors performing separately, overall performance was very fast. The *University of Modena / University of Central Florida* team used frame-frame distances computed based on pixels, and based on histograms. They examined frame difference behaviors over time to see if it corresponded to a linear transformation. The system was not optimized for speed.

University of Iowa's system built on previous years' with a cut detection followed by gradual transition detection. Frame similarities were computed based on color histograms, on aggregated pixel distances, and on edges. There are still some issues of combining gradual transition and cut detection logic. The approach taken by the *University of Marburg* was based on frame similarities measured by motion-compensated pixel differences and histogram differences for several frame distances. An unsupervised ensemble of classifiers was then used. SVM classifiers were trained on 2004 data. Performance was good and quite efficient.

At the *University of Rey Juan Carlos* the team concentrated on cut detection by shape and by a combination of shape and color features. Shape used Zernike moments; color used histograms from last year. Combination methods used various logical combinations. The system did well on precision for cuts. The *University of Sao Paolo* approach appears to be fast but not yet among the best. No details on the system were provided to date.

Details from *Florida International University* were not available for this overview.

3.4 Results

As illustrated in Figure 1 and Figure 2, performance on gradual transitions lags, as expected, behind that on abrupt transitions, where for some uses the problem may be considered a solved one. While progress in detection of gradual transitions may be possible, it is not clear what user/application would require such improvement.

Figure 4 depicts the mean runtime in seconds for each system. It should also be noted that some systems may not have been designed for speed. Where available, this information did illuminate systems from a new angle - one that may be critical to some applications but not others.

Although some groups re-used their work from previous years in most cases this was modified or extended in some way, for example the submissions from Tsinghua University, University of Iowa, RMIT University and IBM Research. Two exceptions were the submissions from Imperial College and from CLIPS who indicated they used their 2004 systems on 2005 data, untouched. It is thus interesting to compare the relative performances of these two groups in 2004 and in 2005 as an indicator of how different the tasks in each year were, relative to each other. On examining the performances of these groups in 2004 and 2005 we find that it is very difficult to separate overall performance figures. The submitted runs from both sites in 2005 are better than 2004 in terms of frame preci-

Figure 1: Precision and recall for cuts

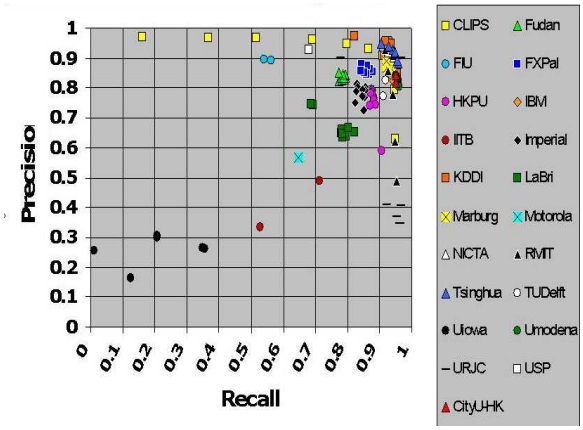


Figure 2: Precision and recall for gradual transitions

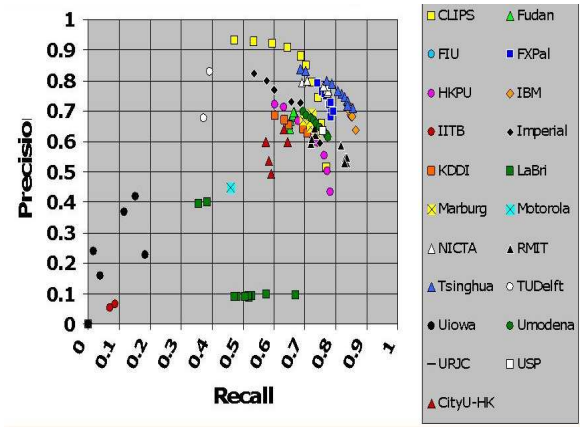


Figure 3: Frame-precision and frame-recall for gradual transitions

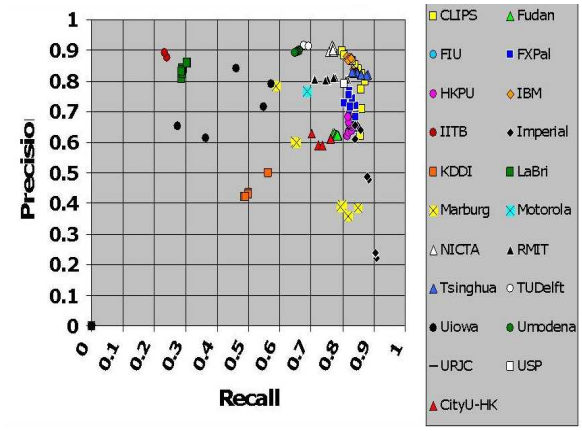
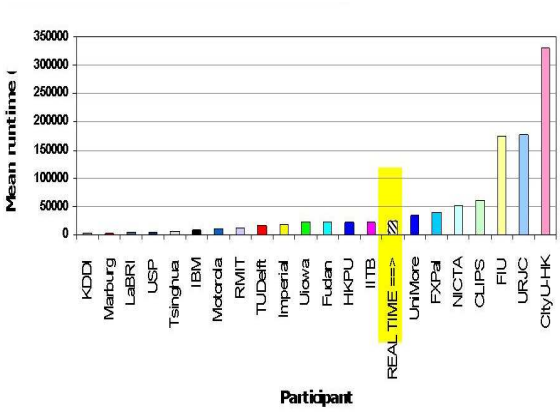


Figure 4: Mean runtime in seconds



sion and frame recall but are identical or 2004 slightly better when we consider overall precision and recall. For hard cuts the CLIPS site is marginally better in 2005 than its own submissions in 2004 while the reverse is true for Imperial College. In summary we can say that the differences between the two normalizing group performances in 2004 and in 2005 are negligible, indicating that the level of difficulty of the task across the two years is approximately the same.

3.5 Issues

According to the guidelines since 2003, shot boundary evaluation treats short ($<$ 5 seconds) gradual transitions as cuts, whether they occur in the reference or the submission. Some participants have objected to this convention, which TRECVID carries over from an earlier shot boundary evaluation. Experiments on the 2005 submissions show reducing the threshold to 4,3,2, or 1 second has varying effects on some but not all submissions. This issue should be investigated further.

4 Low-level (camera motion) feature extraction

In 2005 TRECVID ran a pilot task aimed at evaluating systems' ability to detect a class of low-level features: camera motion. Queries against video archives for footage to be reused can specify particular views, e.g., panning from the left, zooming in, etc. Although tests have been run on small amounts of constructed data (Ewerth, Schwalb, Tessmann, & Freisleben, 2004), and sports video with restricted camera movement (Tan, Saur, Kulkarni, & Ramadge, 2000), we are not aware of large-scale testing on news video.

TRECVID defined three feature groups though in what follows we may refer to the group by the first feature listed for the group below:

- 1 - pan (left or right) or track
- 2 - tilt (up or down) or boom
- 3 - zoom (in or out) or dolly

The grouping acknowledges the difficulty of distinguishing translation along the x-axis (pan) from rotation about the y-axis, etc., and reduced NIST's annotation effort by not requiring the distinguishing of directions (up, down, left, right).

The camera motion task was as follows: given the feature test set, the set of master shot definitions for that test set, and the camera motion feature definitions, return for each of the camera motion features a list of all master shots for which the feature is true. A feature (group) is considered present if it (one or more of its members) occurs anytime within the shot.

4.1 Data

The camera motion task used the same test data as the high-level feature and search tasks. NIST did not provide any training data for the camera motion task. Werner Bailer at Joanneum Research organized a collaborative effort to create such development data using a tool he developed.

4.2 Evaluation

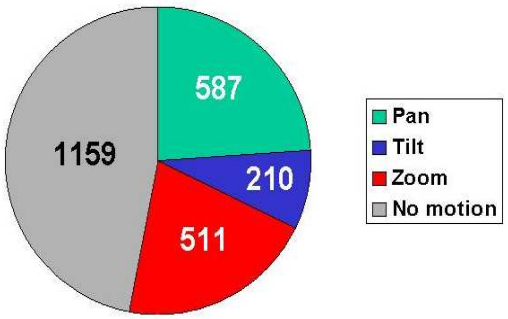
Because the low-level camera movement features are very frequent and often (especially in combination) very difficult even for a human to detect, the low-level feature task was evaluated differently from the high-level feature task.

In advance of any submissions, NIST outlined the procedure to be used in creating the truth data. NIST chose a random subset of the test collection and manually annotate each shot for each of the features. The number of shots was as large as our resources allowed. We allowed ourselves to drop from the annotated subset, shots for which the feature was not clearly true or false in the judgment of the annotator. For example, when a hand-held camera resulted in a minor camera movement in many directions we normally dropped that shot. This was partly to assure that annotations are reliable and because we do not think a user asking, for example, for a panning or tracking shot would want such shaky shots returned.

As it ended up, we had time to look at 5000 shots. From this first pass we kept what seemed reasonably clear examples of each feature (group) as well as examples of shots with no camera motion.

In second pass we doublechecked and corrected the output of the first pass. The ground truth for each feature then consisted of the shots we found for which the feature (group) was true (pan:587, tilt:210, zoom:511) plus the shots we found for which the feature was clearly not true (i.e., the "no motion" shots:1159). See Figure 5. The total number of unique shots is 2226, which amounts to about 4.8 hours of video. In the test subset 844 shots represent just one feature (pan:401,

Figure 5: Motion types found



tilt:92, zoom:351), 205 shots exactly two features (pan/tilt:63, pan/zoom:105, tilt/zoom:37), and 18 shots all three features. The test subset is clearly not a simple random sample and we have not attempted to balance the relative size of any of the sets.

The test subset from each submitted run was then evaluated against the truth data using a script created by NIST and made available to participants.

NIST created three automatic baselines runs:

- Assert feature is true for every shot
- Assert feature is true for a randomly selected subset of the test set, where the subset contains just as many true shots for that feature as the truth data do.
- Choose feature true/false randomly for each shot

4.3 Measures

Each run was evaluated and the basic agreement between the submission and the ground truth was reported in terms of the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). In addition precision [TP/(TP+FP)] and recall [TP/(TP+FN)] and their means (over all three features) were calculated for each run.

4.4 Approaches in brief

Carnegie Mellon University used a probabilistic model (fitted using EM) based on MPEG motion vectors. They also implemented an optical flow model by extracting the most consistent motion from the frame to frame optical flows. The team at *City University of Hong Kong* extracted motion features by tracking image features in consecutive frames. They estimated a 6 parameter affine model with transformation into p,t,z vector for each set of adjacent frames. Their system included rule-based motion classification using empirical thresholds and they performed some interesting failure analysis. *Fudan University* extracted motion vectors from MPEG. They used SVM and a motion accumulation method to filter out imperceptable movements.

Researchers at the *Institute for InfoComm Research* annotated 24 video files. They estimated an affine camera model based on MPEG motion vectors, transformed the parameters into a series of p,t,z values for each shot, and used rule-based classification of series using accumulation and thresholding. At *Joanneum Research* they developed a training set using their annotation tool. Using the training data, they built a system incorporating feature tracking, clustering trajectories, selection of dominant clusters, camera motion detection, and thesholding. *LaBRI* at the University of Bordeaux used MPEG motion vector input to build a 6 parameter affine model. They incorporated Jitter suppression (statistical significance test), subshot segmentation (homogeneous motion), and motion classification (using "a few annotated videos").

MediaMill (University of Amsterdam) started from an existing system based on spatiotemporal image analysis and experimented with modifications such as use of a tesselation of 8 regions on each input frame to reduce the effect of local disturbances, early versus late fusion, and the use of the concept lexicon. Results suffered from a conservative base detector but the use of region-based detectors looked promising. *Tsinghua University's* system employed motion vector selection-based spatial features, separating camera motion from object motion and accidental motion, a 4-parameter camera model (Iterative Least Squares) parameter estimation, and rule-based classification (FSA), using a range of thresholds for: 1. continuous (speed) and noticeable, 2. minimum duration, 3. uninterrupted, 4. noticeable in case of combination with other camera movements.

University of Central Florida based their approach on the analysis of the homography transformation and the fundamental matrix between two consecutive frames. *University of Iowa's* system employed a sliding region window with pixel distance similarity aggregated with a run length threshold. The number of frames in the runlength and the number of

Figure 6: Mean precision and recall by system

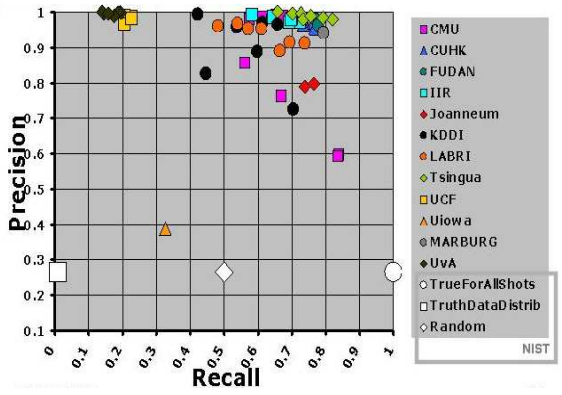


Figure 7: Pan precision and recall by system

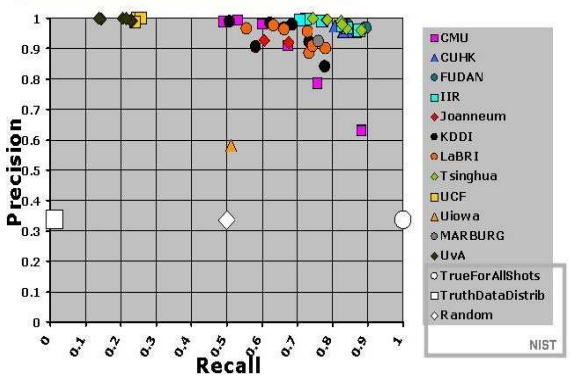


Figure 8: Tilt precision and recall by system

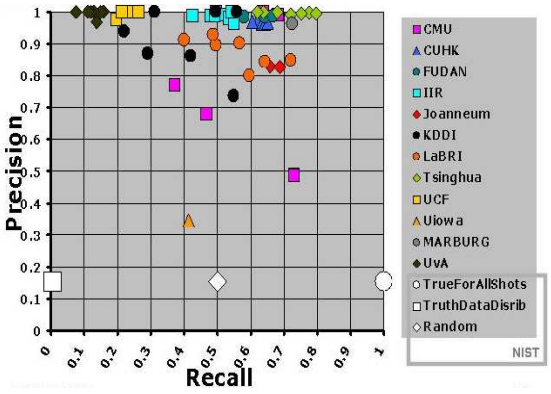
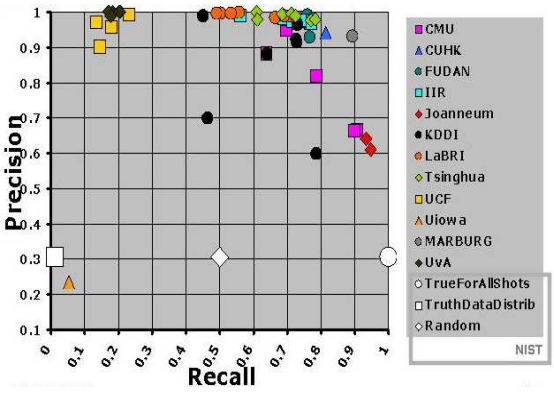


Figure 9: Zoom precision and recall by system



pixels in the window range were varied with no distinction in performance as evaluated. The *University of Marburg* used a $3D$ camera model estimated from MPEG motion vectors from P-frames. Some cleaning was necessary as was exclusion of the center and frame border. Optimal thresholds were estimated on the collaborative TRECVID 2005 training set.

Details from *Bilkent University* and *National ICT Australia* were not available for this overview.

4.5 Results

Information on results is depicted in Figures 6, 7, 8, and 9.

White elements (square, diamond and circle) represent the three automatic NIST baseline runs as explained in subsection 4.2. We opted not to use the obvious *accuracy* measure for evaluation because it

conveys the right intuition only when the positive and negative populations are roughly equal in size. Recall and precision together form a better measure BUT what to do when A has better recall than B and B has better precision than A is not clear The most common approach in this case would be to compute the F-measure (harmonic mean of recall and precision) but for our task this would be misleading. The greater clarity of no-motion shots in the test set should make false positives less likely than false negatives and higher precision easier to achieve than higher recall. So, the farther to the upper right corner the results are from the baseline NIST runs, giving more weight to higher recall, the better the ability to detect camera motion is.

Participants' approaches vary but many of them extract motion vectors directly from compressed video data rather than use optical flow. Some of them tried both and obtained higher results when using MPEG motion vectors. Participants' results show that probabilistic approaches can be used to obtain high recall when detecting low level camera motion. To filter out imperceptible movements and classify camera motion, participants used different learning techniques. It seems that SVMs can classify camera motion more accurately and efficiently than other techniques such as rule-based decision trees. Some groups set a fixed threshold in their systems and only return the shots with high confidence, thus they obtain a bigger precision/recall ratio.

For this task one of the main problems turns out to be the distinction of camera motion from object motion. Best results were achieved with approaches with well defined features and rules, estimation of affine model parameters of camera motion and SVM based classification.

Main results conclusion is that participants obtain higher results for pan, followed by zoom then tilt. We consider that the difficulty in achieving higher recall for tilt is logical. The outliers on the bottom of the video shots can easily misclassified them as pan.

4.6 Issues

The difficulties involved in creating the truth data meant that the test set was not as large as desired. Also, the method does not yield a simple random sample of the test set so that generalization to the entire test set is not simple.

5 High-level feature extraction

A potentially important asset to help video search/navigation is the ability to automatically identify the occurrence of various semantic features such as "Indoor/Outdoor","People", "Speech" etc., which occur frequently in video information. The ability to detect features is an interesting challenge by itself but it would take on added importance if it could serve as a reusable, extensible basis for query formation and search. The feature extraction task has the following objectives:

- to continue work on a benchmark for evaluating the effectiveness of detection methods for various semantic concepts
- to allow exchange of feature detection output for use in the TRECVID search test set prior to the search task results submission date, so that a greater number of participants could explore innovative ways of leveraging those detectors in answering the search task queries in their own systems.

The feature extraction task was as follows. Given a standard set of shot boundaries for the feature extraction test collection and a list of feature definitions, participants were asked to return for each feature that they chose, at most the top 2,000 video shots from the standard set, ranked according to the system's confidence about the feature being present for the shot concerned. During human assessment of the pooled submissions, the presence of each feature was assumed to be binary, i.e., it was either present or absent in the given standard video shot. If the feature was true for some frame (sequence) within the shot, then it was true for the shot. This is a simplification adopted for the benefits it afforded in pooling of results and approximating the basis for calculating recall.

The feature set was a subset of a preliminary set of features developed within the framework of the ARDA/NRRC workshop on Large Scale Ontology for Multimedia (LSCOM), chosen to cover a variety of target types (people, things, locations and activities). It was chosen before the number of instances in the development data was known.

The number of features to be detected was kept small (10) so as to be manageable in this iteration of TRECVID and the features were ones for which more than a few groups could create detectors. Another consideration was whether the features could,

in theory at least, be used in executing searches on the video data as part of the search task, though the topics did not exist at the time the features were defined. Finally, feature definitions were to be in terms a human judge could understand. Some participating groups made their feature detection output available to participants in the search task which really helped the search task and contributed to the collaborative nature of TRECVID.

The features to be detected were defined (briefly) as follows and are numbered 38-47: [38] People walking/running, [39] Explosion or fire, [40] Map, [41] US flag, [42] Building exterior, [43] Waterscape/waterfront, [44] Mountain, [45] Prisoner, [46] Sports, [47] Car. Several have been used before or are similar to previously used ones. The full definitions provided to system developers and NIST assessors are listed in Appendix 9.

5.1 Data

As mentioned above, the feature test collection contained 140 files/videos and 45,765 reference shots. Testing feature extraction and search on the same data offered the opportunity to assess the quality of features being used in search. Training data was available for participants in the collaborative common feature annotation effort (cf. section 2.3).

5.2 Evaluation

Each group was allowed to submit up to 7 runs. In fact 22 groups submitted a total of 110 runs. This is a significant increase with respect to 2004, when only 12 groups participated. Almost all groups submitted runs for all features. Each run had to be annotated with the type of training data set used (cf. section 2.3). Most groups submitted runs of category A, which increased comparability of results between groups.

All submissions down to a depth of 250 result items (shots) were divided into strata of depth 10. So, for example, stratum A contained result set items 1-10 (those most likely to be true), stratum B items 11-20, etc. A subpool for each stratum was formed from the unique items from that stratum in all submissions and then randomized. Assessors were presented with the subpools in "alphabetical" order until they judged all the subpools or ran out of time. The maximum result set depth judged and pooling and judging information for each feature is listed in Table 3. In all, 76,116 shots were judged. The percentage of judged shots that was true ranged between 0.8% and 45.8%. This means that for a few of the features, the 2005 HLF test collection is less very reliable for the evaluation of new experiments, since there are many true shots that have not been judged.

5.3 Measures

The trec_eval software, a tool used in the main TREC activity since it started in 1991, was used to calculate recall, precision, average precision, etc., for each result. In experimental terms the features represent fixed rather than random factors, i.e., we were interested at this point in each feature rather than in the set of features as a random sample of some population of features. For this reason and because different groups could work on very different numbers of features, we did not aggregate measures at the run-level in the results presentations. Comparison of systems should thus be "within feature". Note, that if the total number of shots found for which a feature was true (across all submissions) exceeded the maximum result size (2,000), average precision was calculated by dividing the summed precisions by 2,000 rather than by the the total number of true shots.

5.4 Approaches in brief

Carnegie Mellon University tested unimodal versus multimodal approaches as in 2004. Their system learned dependencies between semantic features (by using various graphical model representations) though results were inconclusive. They found local fusion outperformed global fusion, multilingual outperformed monolingual runs, and multiple text sources proved superior to single text sources. *CLIPS-LSR-NII* explored the use of a 3-level network of stacked classifiers based only on visual information. The objective of this architecture was to leverage contextual information at various level of the analysis process. Results showed that the contextual approach outperformed the baseline approach for all features. The researchers at *Columbia University* experimented with a parts-based object representation that captures topological structure (spatial relationships among parts) and the local attributes of parts. The model learns the parameter distribution properties due to differences in photometric conditions and geometry. Experiments showed that the parts-based approach is indeed an effective approach, improving over a strong baseline by about 10%. The approach seems especially powerful for detecting features that

can be characterized by local attributes and topology, such as "US-flag".

Fudan University submitted several runs: with specific feature detectors, using ASR, and fusing several unimodal SVM classifiers. They ran contrastive experiments with different dimension reduction techniques (e.g., PCA, locality preserving projection). Experiments showed that there was no significant difference between the dimension reduction techniques, but that dimension reduction in itself is an effective technique. The *FX Palo Alto Laboratory* team trained an SVM on low-level features donated by CMU and explored classifier combination schemes based on various forms of regression. The *Helsinki University of Technology's* system was based on self-organizing maps trained on multimodal features and LSCOM lite annotations. *IBM* carried out experiments in fusion across features and across approaches in a flat as well as hierarchical manner. They used support vector machines for learning low-level visual, textual, and meta-features (channel, time, language). They also built models for some features using a modified nearest neighbor learner, a maximum entropy learner, and a Gaussian mixture model. For some regional features a new generalized multiple instance learning algorithm was used. Results indicated both hierarchical feature fusion and fusion across approaches are effective techniques.

Imperial College London worked on "naive" models, locating salient clusters in feature space and learning correspondences between high-level features and the clusters. They also evaluated an approach based on nonparametric density estimation (kernel smoothing). The latter model achieved competitive performance. *Institute Eurecom* compared fusion methods based on support vector machines, with fusion based on hidden Markov models (HMM), and one which fused the SVM and HMM results (using genetic algorithms or SVM). The hierarchical fusion method using genetic algrorithms performed at about median participant level. *Johns Hopkins University* investigated the use of HMMs extended to handle visual and textual features of keyframe images. They combined the posterior probability vectors produced by the HMMs using support vector machines to improve detection. *Language Computer Corporation* tested two classification-based approaches. One employed the k nearest neighbor's method (using Euclidean distance similarity) to cluster development shots and to classify test shots based on the keyframe only. The other used only the ASR text to learn feature models.

LIP6 (University of Paris) researchers tested several variant methods based on fuzzy decision trees on feature 40. The *Lowlands team (CWI, University of Twente, University of Amsterdam)* experimented with feature detectors based on visual information only and compared Weibull-based and GMM-based detectors. Success for any given sort of model varied by topic, suggesting some sort of combination might be useful. The *Mediamill team* at the University of Amsterdam continued their experiments based on the authoring metaphor using automatically learned feature-specific combinations of content, style and context analysis, and a 101 concept lexicon. For them textual features contributed only a small performance gain. The *National University of Singapore (NUS)* explored two methods: ranked maximal figure of merit (known from text categorization) and an HMM followed by RankBoost fusion. Best results were achieved with the latter approach. *Tsinghua University's* approaches relied heavily on visual information. They compared the use of regional versus global features using support vector machine classifiers and the Relay Boost algorithm, respectively.

The *University of Central Florida* experimented with 3 approaches. The first was based on global features that were subdivided into fixed-sized patches. The second approach was based on local features of image segments and the third approach used feature points and appearance similarity. *University of Electro-Communications* investigated the extent to which the high-level concepts in TV news video can be detected based on visual knowledge gleaned from weakly annotated images from the WWW. They used a GMM-based generative model trained on Web images, the TRECVID common feature annotation data, or their combination.

Details from *Bilkent University*, *National ICT Australia*, *SCHEMA (University of Bremen)*, and *University of Washington* were not available for this overview.

5.5 Results

Most groups are now building detectors for all the tested features — the trend is toward generic methods for construction of feature detectors. True shots were found across the three language sources as can be see in Figure 13. Absolute scores (see Figure 10) are generally higher than last year but scores cannot in general be compared directly since at least the data are quite different.

Figure 10: Average precision by feature (boxplot)

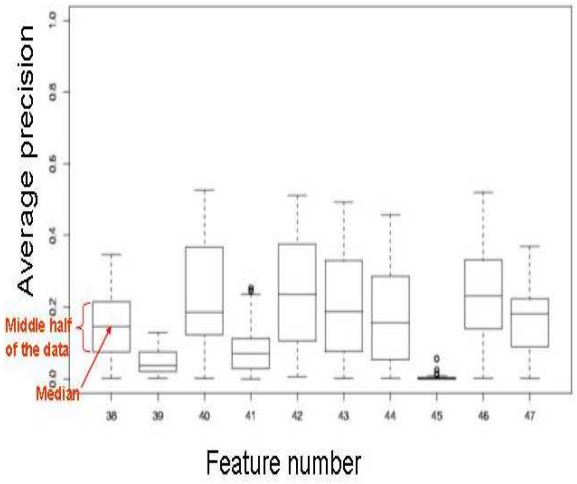


Figure 11: Average precision for top 10 runs

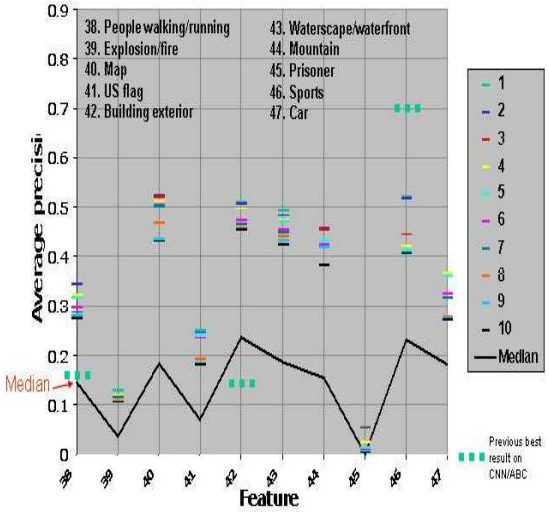
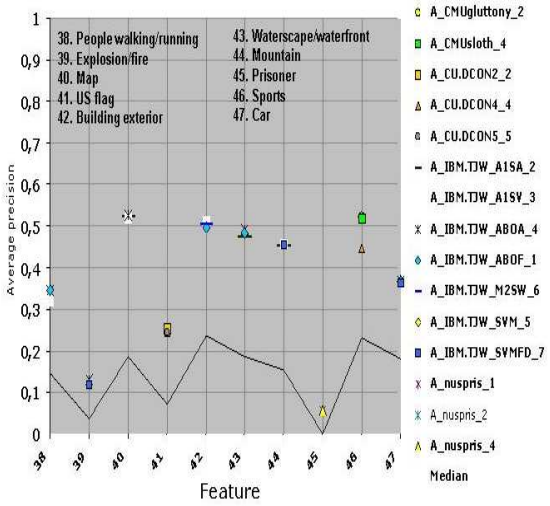


Figure 12 shows top 3 runs per feature when ordered by average precision all from from systems trained only on the common training data (condition A). All of these runs came from only four groups. Figure 11 shows how close together the results for the top ten systems are for most features. Yet some groups' systems have found true shots found by no others, as depicted in Figure 14. Top runs have quite different approaches, but all of them

Figure 12: Average precision for top 3 runs by feature



Conclusions about the relative effectiveness of one approach over another are normally meaningful only within the context of a particular group's experiments, as described in the individual groups' papers on the TRECVID website.

5.6 Issues

The repetition of video material in commercials and in repeated news segments can increase the frequency of true shots for a feature and reduce the usefulness of the recall measure. The extent of this redundancy and its effect on the evaluation have yet to be examined systematically.

The issue of interaction between the feature extraction and the search tasks still needs to be explored so that search can benefit more from feature extraction.

Figure 13: True shots by language and feature

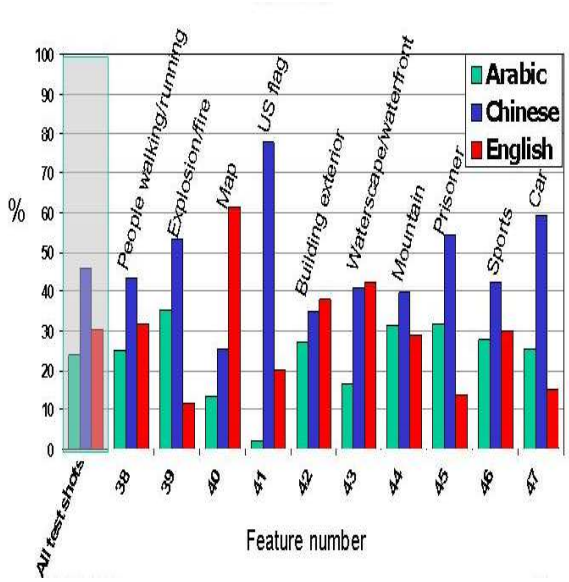
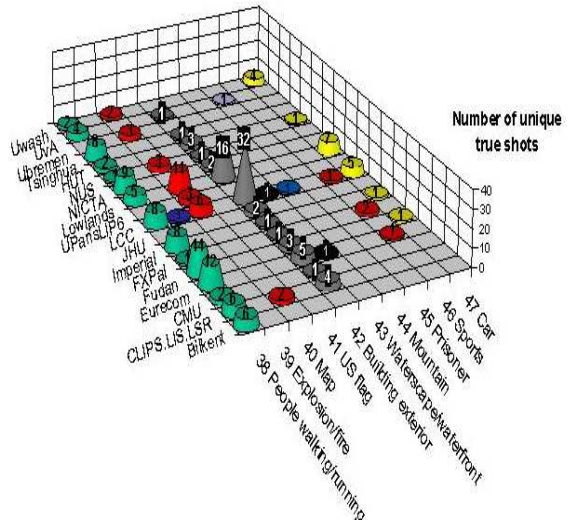


Figure 14: True shots contributed uniquely by team and feature



6 Search

The search task in TRECVID was an extension of its text-only analogue. Video search systems were presented with multimedia topics — formatted descriptions of a need for video — and were asked to return a list of up to 1,000 shots from the videos in the search test collection which met the need. The list was to be prioritized based on likelihood of relevance to the need expressed by the topic.

6.1 Interactive, manual, and automatic search

As was mentioned earlier, three search modes were allowed, fully interactive, manual, and fully automatic. A big problem in video searching is that topics are complex and designating the intended meaning and interrelationships between the various pieces — text, images, video clips, and audio clips — is difficult. The examples of video, audio, etc. do not always represent the information need exclusively and exhaustively. Understanding what an image is of/about is famously complicated (Shatford, 1986).

The definition of the manual mode allowed a human, expert in the search system interface, to interpret the topic and create an optimal query in an attempt to make the problem less intractable. The cost of the manual mode in terms of allowing comparative evaluation is the conflation of searcher and system effects. However if a single searcher is used for all manual searches within a given research group, comparison of searches within that group is still possible. At this stage in the research, the ability of a team to compare variants of their system is arguably more important than the ability to compare across teams, where results are more likely to be confounded by other factors hard to control (e.g. different training resources, different low-level research emphases, etc.).

One baseline run was required of every manual system — a run based only on the text from the provided English ASR/MT output and on the text of the topics. A baseline run was also required of every automatic system — a run based only on the text from the provided English ASR/MT output and on the text of the topics. The reason for the baselines is to help provide a basis for answering the question of how much (if any) using visual information helps over just using text.

6.2 Topics

Because the topics have a huge effect on the results, the topic creation process deserves special attention here. Ideally the topics would have been created by real users against the same collection used to test the systems, but such queries were not available.

Alternatively, interested parties familiar in a general way with the content covered by a test collection could have formulated questions which were then checked against the test collection to see that they were indeed relevant. This is not practical either because it presupposed the existence of the sort of very effective video search tool which participants are working to develop.

What was left was to work backward from the test collection with a number of goals in mind. Rather than attempt to create a representative sample, NIST tried to get an equal number of each of the basic types: generic/specific and person/thing/event, though in no way do we wish to suggest these types are equal as measured by difficulty to systems. Another important consideration was the estimated number of relevant shots and their distribution across the videos. The goals here were as follows:

- For almost all topics, there should be multiple shots that meet the need.
- If possible, relevant shots for a topic should come from more than one video.
- As the search task is already very difficult, we don't want to make the topics too difficult.

The 24 multimedia topics developed by NIST for the search task express the need for video (not just information) concerning people, things, events, locations, etc. and combinations of the former. The topics were designed to reflect many of the various sorts of queries real users pose: requests for video with specific people or types of people, specific objects or instances of object types, specific activities or locations or instances of activity or location types (Enser & Sandom, 2002).

The topics were constructed based on a review of the test collection for relevant shots. The topic creation process was the same as previously – designed to eliminate or reduce tuning of the topic text or examples to the test collection. Potential topic targets were identified watching the test videos with the sound off. Non-text examples were chosen without reference to the relevant shots found. When more examples were found than were to be used, the subset used was chosen at random. The topics are listed in Appendix A. A rough classification of topic types for TRECVID 2005 based on Armitage & Enser, 1996 is provided in Table 5.

6.3 Evaluation

Groups were allowed to submit up to 7 runs. In fact 20 groups (up from 16 in 2004) submitted a total of 112 runs (down from 136) - 44 interactive runs (down from 61), and 26 manual ones (down from 52), and 42 fully automatic ones (up from 23). All 7 runs contributed to the evaluation pools.

All submissions were divided into strata of depth 10. So, for example, stratum A contained result set items 1-10 (those most likely to be true), stratum B items 11-20, etc. A sub-pool for each stratum was formed from the unique items from that stratum in all submissions and then randomized. Assessors were presented with the subpools in "alphabetical" order until they had judged the re-divided set and then ran out of time or stopped finding true shots. At least the top 70 shots were judged completely for each topic. Beyond this, in some cases, the last sub-pool assessed may not have been completely judged. The maximum result set depth judged and pooling and judging information for each feature is listed in Table 4 for details.

6.4 Measures

The trec_eval program was used to calculate recall, precision, average precision, etc.

6.5 Approaches in brief

Carnegie Mellon University participated in the automatic and manual search tasks using a relevance-based probabilistic retrieval model ("ranking logistic regression") to combine diverse knowledge sources. Their system incorporated query typing, query analysis using 14 frequently-used semantic concepts, and 5 types of retrieval components (text, color, texture, edge, and person-X).

Columbia University developed an interactive search tool with text search, CBIR search, story segmentation, story-level browsing, 39 visual concepts from LSCOM-Lite, near-duplicate detection, query-class dependent weights, and cue-X re-ranking. Manual runs used text, CBIR, and visual concepts. Automatic runs used query-class dependent weightings

of some of the above. *Dubin City University* experimented with an interactive search system using a DiamondTouch collaborative tabletop interface from MERL to text and image-based video searching. Two versions were compared: a) one which increases the user's awareness of another user thus forcing the collaboration b) one with "leave me alone" searching support for efficient solo searching. The aim was to explore user-user collaborative search and the findings were that group awareness benefits retrieval. The DCU team also submitted manual and automatic runs – exploring text-only vs. text+image searching;

Fudan University submitted manual runs and explored multi-modal fusion. They found that relation expression fusion was better than linear fusion using a variety of retrieval modalities: text, 14 visual concepts, pseudo relevance feedback, and logistic regression. They also explored training weights online versus training weights offline. The team from *FX Palo Alto Laboratory* participated in interactive search. They enhanced the 2004 system for efficient browsing and enhanced visualization, by adding 29 concepts/semantic features. The system supported story-level browsing, keyframe thumbnails, text dialog overlays, and story timelines; the query comprised text and/or image. Text-only search was as good as text+others (perhaps because the browser and visualization was very strong).

At the *Helsinki University of Technology* a system used for automatic, manual and interactive runs was developed. Experiments addressed text-only vs. text+multi-modal querying. Multi-modal was found to be better than text-only. Interactive search used relevance feedback only with no "search" or shot-level browsing leading to a system with very dynamic user control. The system from *Imperial College London* incorporated content-based search with nearest neighbor browsing in a two-dimensional GUI map browser – an enhancement on their 2004 system with a new kind of relevance feedback. Text-based search, content-based search with relevance feedback and temporal browsing were integrated into one interface with emphasis on supporting the user task.

IBM focused heavily on automatic search. Their automatic system combined speech-based retrieval, visual retrieval using two lightweight learning approaches, and model-based reranking using the 39 concepts from the TRECVID 2005 common annotation effort. The speech-based component included extensive text analysis and 3 kinds of automatic query refinement. The visual component explored a combination of SVMs and a modified nearest neighbor approach (MECBR).

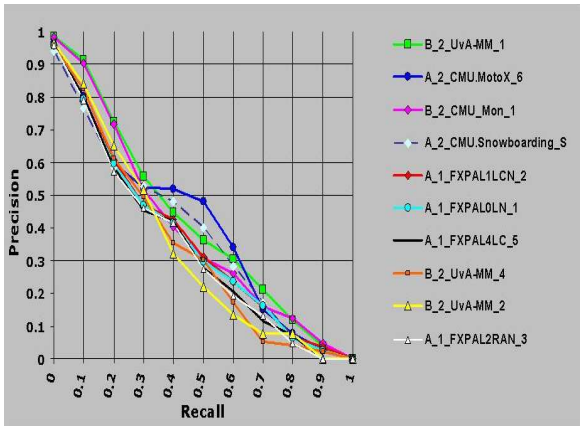
The *Language Computer Corporation* participated in the automatic search task using combinations of ASR text search (language modeling), image features, high-level features, alone and in combination. The image features used blobs. Text search alone was the best-performing which was somewhat unusual in the context of results obtained by other groups. The *Lowlands (CWI, Twente, University of Amsterdam)* team submitted manual and automatic search runs using visual and text searching – first steps towards developing parameterized search engines for each. Weibull and Gaussian mixture models were used for visual features and language modeling for text. In automatic runs and manual runs, using the image in addition to text alone or text and high-level features did not yield a significantly better result.

The *MediaMill (University of Amsterdam and TNO)* team submitted automatic, manual, and interactive search runs using a learned lexicon of 101 semantic concepts and analysis of visual and textual similarity. Automatic runs used only the topic text as input. The manual runs used only the visual modality. In interactive searching various visualizations support visual, hierarchical, and semantic thread browsing.

Researchers at the *National University of Singapore* worked on the automated search task. The test collection was processed to extract text from speech, video OCR, high-level features, audio genre, shot genre, story boundaries, and spatio-temporal information about events. At search time the query was processed to extract keywords, determine query type, event-based modeling, and traditional query expansion. Text from the query is used to retrieve related news articles from the Web. These are used to enhance the query.

Tsinghua University's system supported three search modes - text, image match based on region matching, and concept matching in a concept. The concept/feature recognition approach was based on their HLF submissions. They explored latent relationships (LSA) between (ASR) text and visual features and concepts. They tried each of these alone and in combinations using score fusion and query type-specific weighting. Their conclusion was that combinations worked best. *University of Central Florida* This was UCF's first participation in the search task. Their PEGASUS system, web-based and interactive, used ASR, OCR, keyframe global

Figure 15: Top 10 interactive search runs



histograms and high level features. They submitted ASR-only and multi-modal runs. Multi-modal runs performed better than ASR-only.

The *University of Iowa* submitted automatic runs comparing text-only to text+image features: a) keyframe-keyframe pixel distances; b) text + color information; c) text + texture information; d) text + edge information; they found text-only was best, unlike most other groups. Other combinations would have been possible. The *University of North Carolina at Chapel Hill* investigated the effects of providing context and interactivity in a retrieval system, supporting the browsing of search result sets: a) basic Google-like video search b) enhanced with shot context browsing; c) further enhanced with interactive feedback, e.g., mouseover gives enlarged keyframes; for both performance and user perceptions, the context+interactive system was superior - higher recall, precision the same.

University of Oulu team submitted interactive and manual search runs using a redesigned client application which unites functionality for video queries creation, new cluster-temporal browsing, review of results. The search server formulates subqueries to 3 search subsystems(visual similarity, concepts, and text) and combines the results for presentation to the searcher.

Details from *Bilkent University*,*Queen Mary University of London*, and *SCHEMA - University of Bremen* were not available for this overview.

Figure 16: Top 10 manual search runs

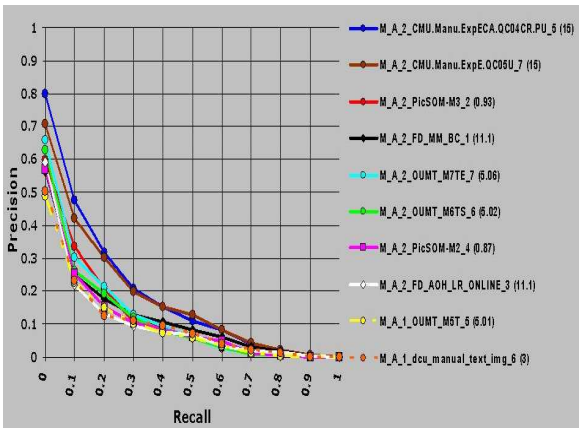
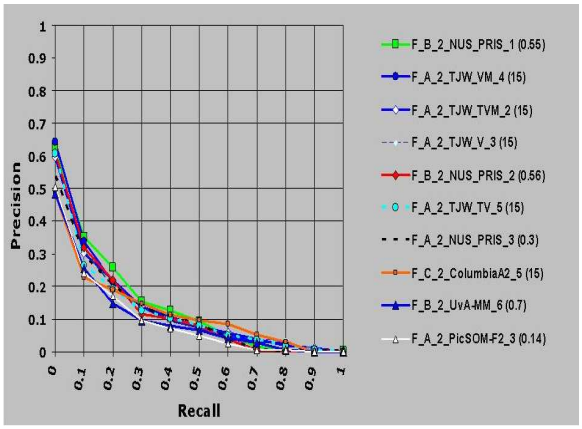


Figure 17: Top 10 automatic search runs



6.6 Results

The 2005 search task introduced some new complexities over previous years, most notably the fact that English speech transcripts were more errorful because the speech in some of the video was Chinese or Arabic. The errors this came from combination of speech recognition and machine translation. Unfortunately, unlike the shot boundary detection task, there were no runs submitted in 2005 which used the same system as used in 2004, so it is not possible to do a direct comparison between years and to measure the effect of the noisy ASR/MT directly.

The results in terms of recall and precision for the top ten interactive, manual, and automatic runs (sorted by mean average precision (MAP)), are presented in Figures 15, 16, and 17 respectively.

¿From these results we can see that the errorful ASR and added noise from machine translation did not prevent systems from finding video that met the needs described in the topics though it did mean that some groups (IBM Research and MediaMill on 16 of the 24 topics) found their visual-only search performed better than their text-only. This indicates that groups are improving the ways in which visual search is being used. Most groups did use both the text and the visual examples in the topic definitions, usually in some multimodal combination. Multimodal approaches have always been common in TRECVID, specifically combinations of retrieval based on searching the ASR text, based on matching keyframes using image similarity approaches, and based on using automatically-derived features. Results from the runs in 2005 showed that multimodal approaches were usually better than unimodal ones and, as might be expected, the visual modality may have been more useful than in previous years.

Beyond that, the conclusions reached by the participants tended to be quite narrow and focused on their own system configurations and on issues they chose to investigate directly.

While there are many variables across sites in the interactive search task, automatic runs can be compared across sites. Among the top 10 automatic runs when ranked by MAP, and using only the common training data, a partial pairwise randomization test (Manly, 1997) on the difference in mean average precision scores shows F_A_2_TJW_TVM_2 to be significantly better than F_A_2_PicSOM-F2 (p=0.029) and F_A_2_TJW_TV_5 (p=0.043). It shows F_A_2_TJW_VM_4 to be better than F_A_2_TJW_V_3 (p=0.015).

When we compare manual runs across sites, we are comparing not just systems but searcher-system pairs. The top 10 manual runs when ranked by MAP are all trained only on the common training data. A partial pairwise randomization test on the different in MAP ($p_¡=0.05$) shows M_A_2_CMU.Manu.ExpECA.QC04CR.PU_5 to have performed better than 7 other runs, M_A_2_CMU.Manu.ExpE.QC05U_7 better than 4 others, and M_A_2_PicSOM-M3_2 better than 1 other. Issues with experimental design make comparison of interactive runs across sites especially problematic. The TRECVID website's tools link has more information on the randomization test used.

Figure 18 shows the number of relevant shots found uniquely by one given site. These provide information about the usefulness of the truth data had the site not contributed to the judged pools, e.g., had the site not participated in TRECVID 2005 but wanted to use the truth data later. The numbers of unique are generally small relative to the total relevant for a given topic, but further analysis is needed to draw strong conclusions.

Figure 19 shows the variation in precision by topic. This reveals quite a lot of variation in the difficulty associated with different topics with some topics (tennis player and soccer match goal for example) demonstrating quite good retrieval performance and others (people entering/leaving a building) proving to be very difficult. Figure 20 shows the median average precision across systems by topic for interactive, manual, and automatic runs and the large variation in performance can clearly be seen in these graphs.

In this overview we have been able to present only a small amount of the analysis of results which the large effort participants have put into the search task, deserves. Further analysis should be carried out to try to answer other outstanding questions. For example figure 21 shows the effect of training type (A = common training data only, B = other) for runs using text plus other information. There are in general many more A runs than B.

Figure 22 shows the effect of condition (1 = text only, 2 = other) for runs from systems trained only on the shared training data. There are in general many more condition 2 runs than condition 1. Figure 23 also shows the effect of using more than text in the search but does so by group, where runs are more comparable.

Figure 19: Mean average precision by topic

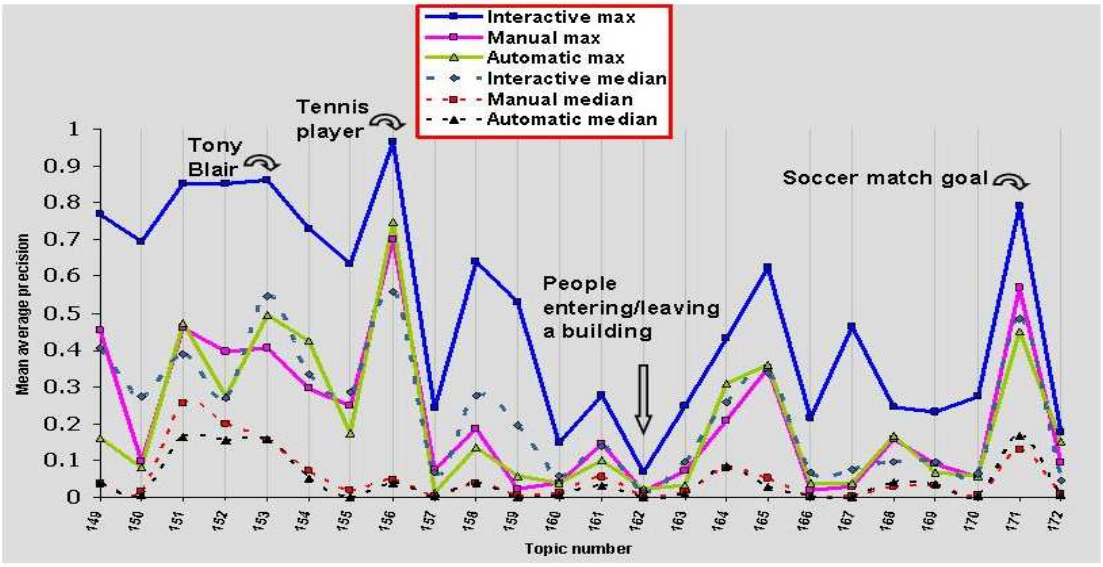


Figure 23: Effect of condition (1=text only, 2=other) for training type A runs by group

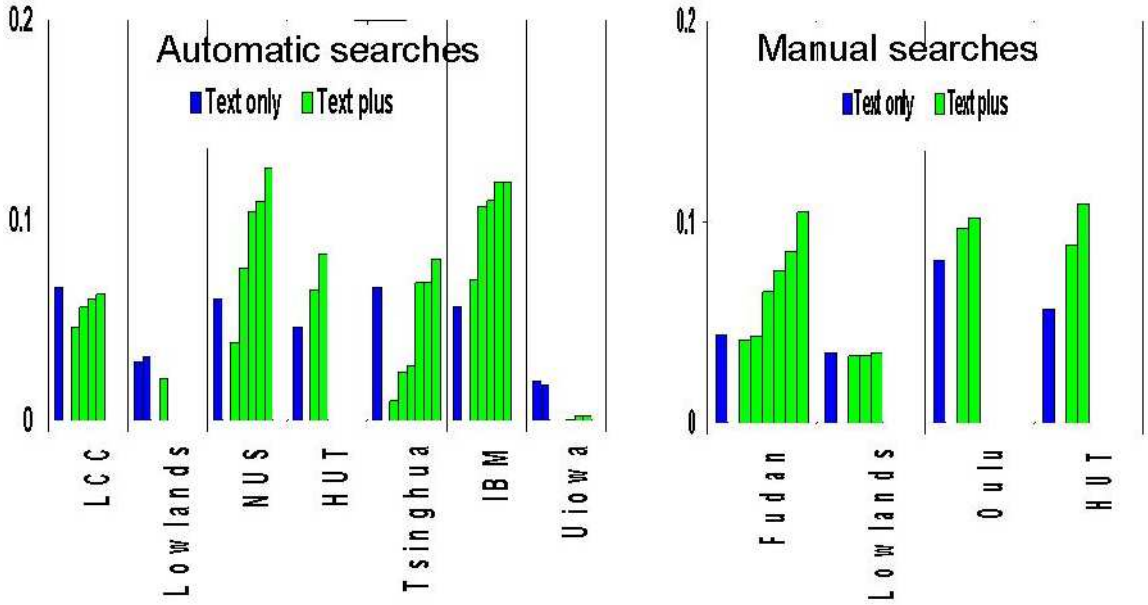


Figure 18: Relevant shots contributed uniquely by group and topic

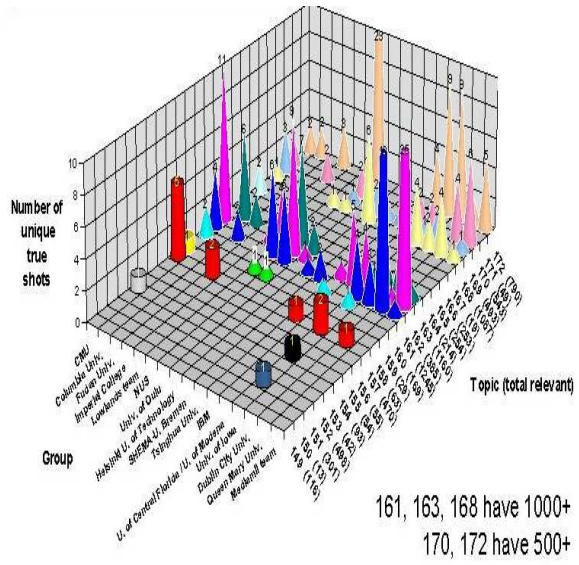


Figure 20: Topics sorted by median mean average precision

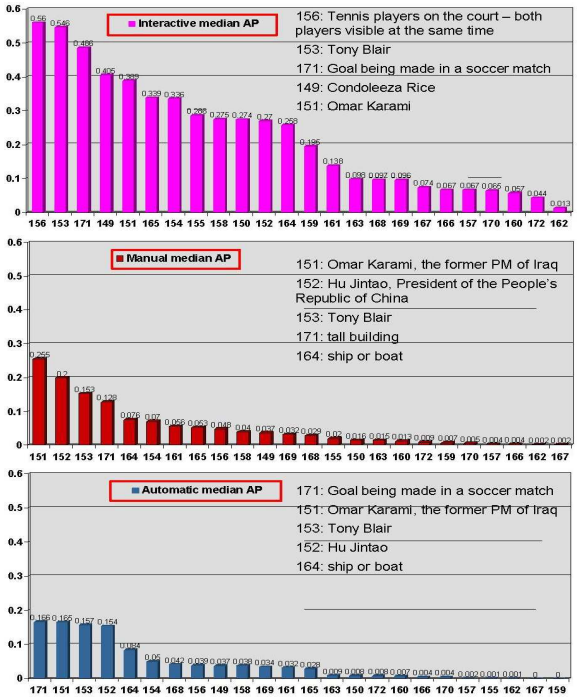


Figure 21: Effect of training type (A=common training data, B=other) for condition 2 (text+x) runs

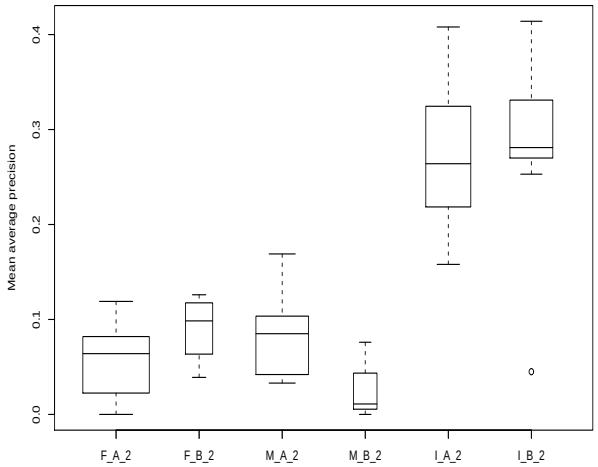
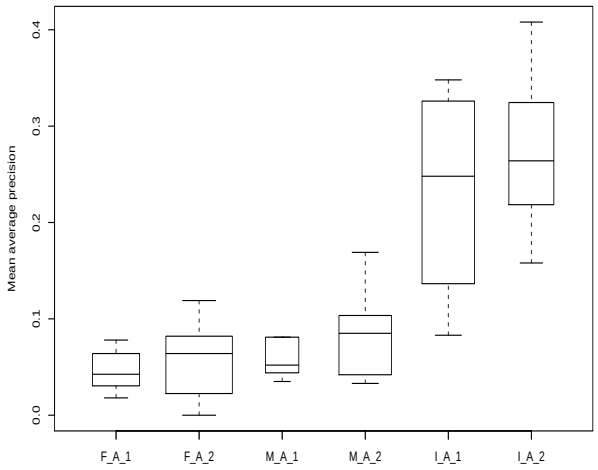


Figure 22: Effect of condition (1=text only, 2=other) for training type A runs



6.7 Issues

7 BBC rushes management

Rushes are the raw video material used to produce a video. Twenty to forty times as much material may be shot as actually becomes part of the finished product. Rushes usually have only natural sound. Actors are only sometimes present. Rushes contain many frames or sequences of frames that are highly repetitive, e.g., many takes of the same scene redone due to errors (e.g. an actor gets his lines wrong, a plane flies overhead, etc.), long segments in which the camera is fixed on a given scene or barely moving, etc. A significant part of the material might qualify as stock footage - reusable shots of people, objects, events, locations. Rushes are potentially very valuable but are largely unexploited because only the original production team knows what the rushes contain and access is generally very limited, e.g., indexing by program, department, name, date (Wright, 2005).

The BBC Archive provided about 50 hours of rushes shot for BBC travel programming along with some metadata and keyframes created by a proprietary asset management system. TRECVID participants were invited to 1) build a system to help a person, unfamiliar with the rushes browse, search, classify, summarize, etc. the material in the archive. 2) devise their own way of evaluating such a system's effectiveness and usability.

7.1 Approaches

Accenture Technology Labs and Siderean Software developed a system using both the textual metadata (including subject description keywords) provided and MPEG-7 low-level visual, color, and textual features they extracted from the provided keyframes. Where possible, subject description terms were linked to concepts in the Library of Congress's Thesaurus of Graphical Materials. The user interface allowed for navigation over the shot database using facets derived from textual and visual metadata.

City University of Hong Kong experimented with methods for structuring and characterizing video content by using motion to infer intention. Their intuition was that such information should eventually be helpful for search, browsing, and summarization.

Dublin City University looked at the utility of letting the searcher use video objects in place of or in addition to whole keyframes in the search process. They constructed and compared two corresponding systems.

IBM examined the applicability of existing semantic models from other domains (news, personal photo annotations) when applied to the rushes video and found many concepts with consistent definitions across domains, but also a few production-specific concepts and surprising re-definitions. They also looked at building a higher-level pattern discovery capability on top of a large lexicon (LSCOM) of concepts and found expected patterns (water-outdoors) as well as novel ones (studio-person : people dancing in a nightclub).

The *Mediamill (University of Amsterdam, TNO)* team evaluated support vector machine models, which had been trained on TRECVID news data, against the BBC rushes. They found 25 of the 39 concepts "survived" – evidence for cross-domain usability.

University of Central Florida investigated a rushes management system eventually to be a content-based image retrieval system, where the content is based on the indexing of the interest points rather than traditional region features.

The most obvious outcome from the BBC rushes task this year was to show that the groups who took part developed very different approaches to rushes management. Also, as a "pre-track',' including the BBC rushes exploration activity in 2005 showed that there are several groups willing and able to manage this volume of completely unstructured video and the activity in 2005 will help shape the task in 2006 and possibly beyond.

8 Summing up and moving on

This overview of TRECVID 2005 has provided basic information on the goals, data, approaches, evaluation mechanisms/metrics, and results. Further details about each particular group's approach and performance can be found in that group's notebook paper and/or slides in the TRECVID on-line proceedings: www-nlpir.nist.gov/projects/trecvid. The interest in TRECVID and the participation continues to grow stronger each year and we look forward with anticipation to future TRECVIDs.

9 Authors' note

TRECVID would not happen without support from ARDA/DTO and NIST and the research community is very grateful for this. Beyond that, various individuals and groups deserve special thanks.

Richard Wright at the BBC Archive made the rushes data available and Gary Marchionini and the Open Video Project at the University of North Carolina at Chapel Hill helped us get the NASA videos in MPEG-1 format.

We are particularly grateful to Christian Petersohn at the Fraunhofer (Heinrich Hertz) Institute in Berlin for providing the master shot reference and to the team at the Centre for Digital Video Processing at Dublin City University (DCU) for formating the master shot reference definition and selecting keyframes.

DCU, the University of Amsterdam, and the University of Iowa helped out in the distribution of corrected data to replace the corrupted or inaccessible data on the hard drives.

We appreciate Jonathan Lasko's painstaking creation of the shot boundary truth data once again.

Randy Paul was instrumental in arranging for a US government contractor to provide ASR and MT output. Alex Hauptmann and others at Carnegie Mellon University donated ASR and MT output to supplement and complete the initial set.

Timo Volkmer and others at IBM created and supported the use of a new web-based system for collaborative annotation. CMU made their annotation system available.

CMU once again donated a set of features for use by other participants. Columbia University donated story boundaries.

Werner Bailer at Joanneum Research developed a tool for annotation of camera motion and made it available to participants in the low-level feature task.

Finally, we would like to thank all the participants and other contributors on the mailing list for their energy, patience, and continued hard work.

Appendix A: Topics

The text descriptions of the topics are listed below followed in brackets by the associated number of image examples (I), video examples (V), and relevant shots (R) found during manual assessment the pooled runs.

0149 Find shots of Condoleeza Rice (I 3, V 6, R 116)

Table 5: 2005 Topic types

	Named			Generic		
Topic	Person, thing	Event	Place	Person, thing	Event	Place
149	X					
150	X					
151	X					
152	X					
153	X					
154	X					
155				X		
156				X		X
157				X	X	
158				X	X	
159	X			X	X	
160				X	X	
161				X		
162				X	X	
163				X	X	
164				X		
165				X		X
166				X		
167				X	X	
168				X		X
169				X		
170				X		
171					X	
172				X		X

0150 Find shots of Iyad Allawi, the former prime minister of Iraq (I 3, V 6, R 13)

0151 Find shots of Omar Karami, the former prime minister of Lebannon (I 3, V 5, R 301)

0152 Find shots of Hu Jintao, president of the People's Republic of China (I 3, V 9, R 498)

0153 Find shots of Tony Blair (I 3, V 4, R 42)

0154 Find shots of Mahmoud Abbas, also known as Abu Mazen, prime minister of the Palestinian Authority (I 3, V 9, R 93)

0155 Find shots of a graphic map of Iraq, location of Bagdhad marked - not a weather map (I 4, V 10, R 54)

0156 Find shots of tennis players on the court - both players visible at same time (I 2, V 4, R 55)

0157 Find shots of people shaking hands (I 4, V 10, R 470)

0158 Find shots of a helicopter in flight (I 2, V 8, R 63)

0159 Find shots of George Bush entering or leaving a vehicle, e.g., car, van, airplane, helicopter, etc - he and the vehicle both visible at the same time. (I 2, V 7, R 29)

0160 Find shots of something (e.g., vehicle, aircraft, building, etc) on fire with flames and smoke visible (I 2, V 9, R 169)

0161 Find shots of people with banners or signs (I 2, V 6, R 1245)

0162 Find shots of one or more people entering or leaving a building (I 4, V 8, R 385)

0163 Find shots of a meeting with a large table and more than two people (I 2, V 5, R 1160)

0164 Find shots of a ship or boat (I 3, V 7, R 214)

0165 Find shots of basketball players on the court (I 2, V 8, R 254)

0166 Find shots of one or more palm trees (I 2, V 6, R 253)

0167 Find shots of an airplane taking off (I 2, V 5, R 19)

0168 Find shots of a road with one or more cars (I 2, V 5, R 1087)

0169 Find shots of one or more tanks or other military vehicles (I 3, V 8, R 493)

0170 Find shots of a tall building (with more than 5 floors above the ground) (I 2, V 6, R 543)

0171 Find shots of a goal being made in a soccer match (I 1, V 7, R 49)

0172 Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people (I 3, V 8, R 790)

Appendix B: Features

38 People walking/running: segment contains video of more than one person walking or running

39 Explosion or fire: segment contains video of an explosion or fire

40 Map: segment contains video of a map

41 US flag: segment contains video of a US flag

42 Building exterior: segment contains video of the exterior of a building

43 Waterscape/waterfront: segment contains video of a waterscape or waterfront

44 Mountain: segment contains video of a mountain or mountain range with slope(s) visible

45 Prisoner: segment contains video of a captive person, e.g., imprisoned, behind bars, in jail, in handcuffs, etc.

46 Sports: segment contains video of any sport in action

47 Car: segment contains video of an automobile

References

Armitage, L. H., & Enser, P. G. B. (1996). *Information Need in the Visual Document Domain: Report on Project RDD/G/235 to the British Library Research and Innovation Centre.* School of Information Management, University of Brighton.

Enser, P. G. B., & Sandom, C. J. (2002). Retrieval of Archival Moving Imagery — CBIR Outside the Frame. In M. S. Lew, N. Sebe, & J. P. Eakins (Eds.), *Image and Video Retrieval, International Conference, CIVR 2002, London, UK, July 18-19, 2002, Proceedings* (Vol. 2383). Springer.

Ewerth, R., Schwalb, M., Tessmann, P., & Freisleben, B. (2004). Estimation of Arbitrary Camera Motion in MPEG Videos. In *Proceedings of the 17th International Conference on Pattern Recognition* (Vol. I, pp. 512–515). Cambridge, UK.

Lee, A. (2001). *VirtualDub home page.* URL: www.virtualdub.org/index.

Manly, B. F. J. (1997). *Randomization, Bootstrap, and Monte Carlo Methods in Biology* (2nd ed.). London, UK: Chapman & Hall.

Shatford, S. (1986). Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloging and Classification Quarterly, 6*(3), 39—61.

Tan, Y.-P., Saur, D. D., Kulkarni, ., Sanjeev R, & Ramadge, P. J. (2000). Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation. *IEEE Transactions on Circuits and Systems for Video Technology*, *10*(1), 133–146.

Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venogupal, A., Zhao, B., & Waibel, A. (2003). The CMU Statistical Translation System. In *Proceedings of mt summit ix.* New Orleans, LA, USA.

Wright, R. (2005). *Personal communication from Richard Wright, Technology Manager, Projects, BBC Information & Archives.*

Table 1: Participants and tasks

Participants	Country	Task				
Accenture Technology Labs / Siderean Software	USA	–	–	–	–	RU
Bilkent University	Turkey	–	LL	HL	SE	–
Carnegie Mellon University	USA	–	LL	HL	SE	RU
City University of Hong Kong	China	SB	LL	–	–	RU
CLIPS-IMAG, LSR-IMAG, Laboratoire LIS	France	SB	–	HL	–	–
Columbia University	USA	–	–	HL	SE	–
Dublin City University	Ireland	–	–	–	SE	RU
Florida International University	USA	SB	–	–	–	–
Fudan University	China	SB	LL	HL	SE	–
FX Palo Alto Laboratory	USA	SB	–	HL	SE	–
Helsinki University of Technology	Finland	–	–	HL	SE	–
Hong Kong Polytechnic University	China	SB	–	–	–	–
IBM	USA	SB	–	HL	SE	RU
Imperial College London	UK	SB	–	HL	SE	–
Indian Institute of Technology (IIT)	India	SB	–	–	–	–
Institut Eurecom	France	–	–	HL	–	–
Institute for Infocomm Research	Singapore	–	LL	–	–	–
JOANNEUM RESEARCH	Austria	–	LL	–	–	–
Johns Hopkins University	USA	–	–	HL	–	–
KDDI R&D Laboratories, Inc.	Japan	SB	LL	–	–	–
Language Computer Corporation (LCC)	USA	–	–	HL	SE	–
LaBRI	France	SB	LL	–	–	–
LIP6-Laboratoire d'Informatique de Paris 6	France	–	–	HL	–	–
Lowlands Team (CWI, Twente, U. of Amsterdam)	Netherlands	–	–	HL	SE	–
Mediamill Team (Univ. of Amsterdam and TNO)	Netherlands	–	LL	HL	SE	RU
Motorola Multimedia Research Laboratory	USA	SB	–	–	–	–
National ICT Australia	Australia	SB	LL	HL	–	–
National University of Singapore (NUS)	Singapore	–	–	HL	SE	–
Queen Mary University of London	UK	–	–	–	SE	–
RMIT University	Australia	SB	–	–	–	–
SCHEMA-Univ. Bremen Team	EU	–	–	HL	SE	–
Technical University of Delft	Netherlands	SB	–	–	–	–
Tsinghua University	China	SB	LL	HL	SE	–
University of Central Florida / University of Modena	USA,Italy	SB	LL	HL	SE	RU
University of Electro-Communications	Japan	–	–	HL	–	–
University of Iowa	USA	SB	LL	–	SE	–
University of Marburg	Germany	SB	LL	–	–	–
University of North Carolina	USA	–	–	–	SE	–
University of Oulu / MediaTeam	Finland	–	–	–	SE	–
University Rey Juan Carlos	Spain	SB	–	–	–	–
University of Sao Paulo (USP)	Brazil	SB	–	–	–	–
University of Washington	USA	–	–	HL	–	–

Task legend. SB: Shot boundary; LL: Low-level features; HL: High-level features; SE: Search ; RU: BBC rushes

Table 3: Feature pooling and judging statistics

Feature number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number true	% judged that were true
38	176314	33424	19.0	250	9000	26.9	3594	39.9
39	185820	30686	16.5	250	6922	22.6	390	5.6
40	203223	32278	15.9	250	5942	18.4	1995	33.6
41	188162	34834	18.5	250	8956	25.7	522	5.8
42	190673	29281	15.4	250	7639	26.1	3497	45.8
43	194770	30570	15.7	250	6560	21.5	868	13.2
44	194482	31487	16.2	200	7296	23.2	752	10.3
45	180815	38154	21.1	250	10667	28.0	88	0.8
46	178879	31337	17.5	250	6177	19.7	576	9.3
47	186796	29755	15.9	250	6957	23.4	2079	29.9

Table 4: Search pooling and judging statistics

Topic number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number relevant	% judged that were relevant
149	88988	24054	27.0	70	1971	8.2	116	5.9
150	85715	22971	26.8	80	3132	13.6	13	0.4
151	91855	18027	19.6	120	2643	14.7	301	11.4
152	93614	16250	17.4	110	2712	16.7	498	18.4
153	88507	23443	26.5	70	2075	8.9	42	2.0
154	88573	21660	24.5	90	2688	12.4	93	3.5
155	92775	21708	23.4	70	2683	12.4	54	2.0
156	89937	22297	24.8	70	2083	9.3	55	2.6
157	91372	24180	26.5	90	4067	16.8	470	11.6
158	89732	22469	25.0	70	2301	10.2	63	2.7
159	93086	22605	24.3	80	3505	15.5	29	0.8
160	94673	22821	24.1	90	3690	16.2	169	4.6
161	94101	23372	24.8	90	3528	15.1	1245	35.3
162	91813	26796	29.2	110	5934	22.1	385	6.5
163	94181	22324	23.7	120	5072	22.7	1160	22.9
164	89724	22633	25.2	100	2737	12.1	214	7.8
165	90639	21508	23.7	90	2393	11.1	254	10.6
166	92667	25160	27.2	90	3999	15.9	253	6.3
167	87155	23645	27.1	70	2857	12.1	19	0.7
168	91932	20772	22.6	110	3945	19.0	1087	27.6
169	93597	21434	22.9	90	3368	15.7	493	14.6
170	92216	23486	25.5	110	4767	20.3	543	11.4
171	92002	23136	25.1	70	2071	9.0	49	2.4
172	93280	25834	27.7	90	4198	16.2	790	18.8

Table 6: Participants not submitting runs

Participants	Country	Task				
Chinese University of Hong Kong	China	–	–	–	–	–
ETRI (Electronics and Telecommunication Research Institute)	Korea	–	–	–	–	–
Fraunhofer-Institute	Germany	–	–	–	–	–
Indiana University	USA	–	–	–	–	–
Nagoya University	Japan	–	–	–	–	–
National Institute of Informatics	Japan	–	–	–	–	–
National Technical University of Athens (1	Greece	–	–	–	–	–
National Technical University of Athens (2)	Greece	–	–	–	–	–
Oxford University	UK	–	–	–	–	–
Polytechnical University of Valencia	Spain	–	–	–	–	–
Ryerson University	Australia	–	–	–	–	–
SAMOVA Team - IRIT - UPS	France	–	–	–	–	–
Tampere University of Technology	Finland	–	–	–	–	–
University of East Anglia	UK	–	–	–	–	–
University of Geneva	Switzerland	–	–	–	–	–
University of Kentucky	USA	–	–	–	–	–
University of Maryland	USA	–	–	–	–	–
University of Ottawa School	Canada	–	–	–	–	–
University of Wisconsin-Milwaukee	USA	–	–	–	–	–
University of York	UK	–	–	–	–	–

Task legend. SB: Shot boundary; LL: Low-level features; HL: High-level features; SE: Search ; RU: BBC rushes