

TRECVID 2005 - An Introduction

Paul Over {over@nist.gov}
and Tzveta Ianeva {tianeva@nist.gov}
Retrieval Group
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA

Wessel Kraaij {wessel.kraaij@tno.nl}
TNO Information and Communication Technology
Delft, the Netherlands

Alan F. Smeaton {asmeaton@computing.dcu.ie}
Adaptive Information Cluster / Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland

November 28, 2005

1 Introduction

TRECVID 2005 represents the fifth running of a TREC-style video retrieval evaluation, the goal of which remains to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. Over time this effort should yield a better understanding of how systems can effectively accomplish such retrieval and how one can reliably benchmark their performance. TRECVID is funded by ARDA and NIST.

Forty-two teams from various research organizations — 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 4 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 as well as the approaches taken by the participating groups will be presented at the workshop. For detailed information about the approaches and results, the reader should see the various site reports and the results pages at the back of the workshop notebook.

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.

1.1 New in TRECVID 2005

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

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 news video: 43 in Arabic, 52 in Chinese, 74 in English. The 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 collection. It 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.

2.2 Common shot reference, keyframes, ASR

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:

Table 2: News files provided

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

Op-tional	Development				Test			
	MPEG-1	Virage	MS-ASR	XLT of MS-ASR	MPEG-1	Virage	MS-ASR	XLT of MS-ASR
Chi	43		39		42		42	
Eng	68	57	40		68	39	42	

- 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 ¹ VirtualDub was used to view the videos and frame numbers. The distribution of transition types was as follows:

¹The VirtualDub (Lee, 2001) website contains information about VirtualDub tool and the MPEG decoder it uses.

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

See the results pages at the back of notebook for detailed information about the performance of each sub-

mitted run.

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), 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 handheld 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). The total number of unique shots is 2226, which amount to about 4.8 hours of video. In the test subset 844 shots represent just one feature (pan:401, 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 Results

See the tables in the results section of the notebook for details.

4.5 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. 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*? 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 test set makes false positive less likely than false negatives and higher precision easier to achieve than higher recall.

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 highest possibility of detecting the presence of the feature. 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. 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 in 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 with the detailed feature runs at the back of the notebook and in Appendix B.

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.

5.2 Evaluation

Each group was allowed to submit up to 7 runs. In fact 22 groups submitted a total of 110 runs.

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.

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 worked on very different numbers of features, we did not aggregate measures at the run-level in the results pages at the back of the notebook. 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 Results

See the results section at the back of the notebook for details about the performance of each run.

5.5 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.

6 Search

The search task in TRECVID was an extension of its text-only analogue. Video search systems were presented with topics — formatted descriptions of an information need — 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 a complex one and 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 in 2003 – 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 Results

See the results pages at the back of the notebook for information about each search run’s performance.

6.6 Issues

The implications of pooling/judging depth on relevant shots found and on system scoring and ranking have yet to be investigated thoroughly for the current systems and data.

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 over, 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.

Six groups took part in the rushes task. See the site papers in the workshop notebook for details about their approaches and results.

It is hoped that enough will be learned from this exploration to allow the addition of a well-defined task with evaluation in TRECVID 2006.

8 Summing up and moving on

This overview of the TREC-2005 Video Track has provided basic information on the goals, data, evaluation mechanisms and metrics used. Further details about each particular group’s approach and performance can be found in that group’s site report. The raw results for each submitted run can be found in the results section of at the back of the notebook.

9 Authors' note

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

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 forming 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 cover gaps in 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.

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.

10 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)

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 Lebanon (I 3, V 5, R 301)

Table 5: 2005 Topic types

Topic	Named			Generic		
	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

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 Baghdad 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)
- 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

11 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

- 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.
- 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	-	-	HL	SE	-
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	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