Proc. TRECVID 2008 Workshop, November 2008, Gaithersburg, MD, USA.

K-Space at TRECVid 2008

Peter Wilkins, Daragh Byrne, Gareth J.F.Jones, Hyowon Lee,
Gordon Keenan, Kevin McGuinness, Noel E. O'Connor, Neil O'Hare,
Alan F. Smeaton, Tomasz Adamek
Centre for Digital Video Processing & CLARITY: Centre for Sensor Web Technologies
Dublin City University (DCU), Ireland

Raphaël Troncy, Alia Amin
Centrum Wiskunde & Informatica, the Netherlands

Rachid Benmokhtar, Emilie Dumont, Benoit Huet, Bernard Merialdo
Departement Communications Multimedia
Institut Eurécom
2229, route des Crtes, 06904 Sophia-Antipolis, France

Giorgos Tolias, Evaggelos Spyrou, Yannis Avrithis
Image Video and Multimedia Laboratory National Technical University of Athens (NTUA-ITI)
9 Iroon Polytechniou Str., 157 80 Athens, Greece

Georgios Th. Papadopoulous, Vasileios Mezaris, Ioannis Kompatsiaris
Informatics and Telematics Institute (CERTH-ITI)
Thessaloniki, Greece

Roland Mörzinger, Peter Schallauer, Werner Bailer
Institute of Information Systems and Information Management
Joanneum Research (JRS)
Steyrergasse 17, 8010 Graz, Austria

Krishna Chandramouli, Ebroul Izquierdo
Department of Electronic Engineering
Queen Mary, University of London (QMUL), United Kingdom

Lutz Goldmann, Martin Haller, Amjad Samour, Andreas Corbet, Thomas Sikora
Technical University of Berlin, Department of Communication Systems (TUB)
EN 1, Einsteinufer 17, 10587 Berlin, Germany

Pavel Praks
Department of Information and Knowledge Engineering
Faculty of Informatics and Statistics, University of Economics, Prague (UEP)
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic

David Hannah, Martin Halvey, Frank Hopfgartner,
Robert Villa, P. Punitha, Anuj Goyal and Joemon M. Jose
Department of Computing Science, University of Glasgow (UG)
University Avenue, Glasgow G12 8QQ, United Kingdom.

October 28, 2008

Abstract

In this paper we describe K-Space's participation in TRECVid 2008 in the interactive search task. For 2008 the K-Space group performed one of the largest interactive video information retrieval experiments conducted in a laboratory setting. We had three institutions participating in a multi-site multi-system experiment. In total 36 users participated, 12 each from Dublin City University (DCU, Ireland), University of Glasgow (GU, Scotland) and Centrum Wiskunde & Informatica (CWI, the Netherlands). Three user interfaces were developed, two from DCU which were also used in 2007 as well as an interface from GU. All interfaces leveraged the same search service. Using a latin squares arrangement, each user conducted 12 topics, leading in total to 6 runs per site, 18 in total. We officially submitted for evaluation 3 of these runs to NIST with an additional expert run using a 4th system. Our submitted runs performed around the median. In this paper we will present an overview of the search system utilized, the experimental setup and a preliminary analysis of our results.

1 Overview of K-Space

K-Space is a European Network of Excellence (NoE) in semantic inference for semi-automatic annotation and retrieval of multimedia content [1] which is in the second year of its three year funding. It is coordinated by Queen Mary University of London (QMUL) and the partner responsible for coordinating the K-Space participation in TRECVid is Dublin City University. K-Space is focused on the research and convergence of three themes: content-based multimedia analysis, knowledge extraction and semantic multimedia.

2 Search Experiment Overview

As stated in the abstract, our participation in TRECVid 2008 interactive search was to conduct one of the largest interactive video information retrieval experiments in a laboratory setting to date. Our motivation for this was to conduct an investigation into interactive multimedia retrieval which seeked to tease apart as many influencing variables as possible. This paper will primarily detail the systems used in the experiment, our experimental parameters and an initial examination of the results.

3 Common Search Engine

The three user interfaces developed for the search experiment leveraged a common search engine. Components of this common engine leveraged previous content-analysis techniques of K-Space partners that were used in TRECVid 2007. The following briefly details these components, a more complete explanation of these components can be found in last years TRECVid publication [23].

As no common keyframe set was released by TRECVid we extracted our own set of keyframes. Our keyframe selection strategy was to extract every second I-Frame from each shot. We extracted low-level visual features from K-frames using several feature descriptors based on the MPEG-7 XM. These descriptors were implemented as part of the aceToolbox, a toolbox of low-level audio and visual analysis tools developed as part of our participation in the EU aceMedia project. We made use of six different global visual descriptors. These descriptors were Colour Layout, Colour Moments, Colour Structure, Homogeneous Texture, Edge Histogram and Scalable Colour. A complete description of each of these descriptors can be found in [14]. We also segmented the keyframes and extracted region based features. This processing was made available to all K-Space partners, further details available in last year's paper [23].

3.1 Institute EURÉCOM

The Eurecom system for this year employs the approaches used in KSpace TRECVid 2007 HLFE task [4][23], for 36 semantic concepts of the 2008 test collection, in the goal of an incorporatation into the DCU multi-site search system. The Eurecom approach is based on a multi-descriptor system. These descriptors are introduced in separate SVM classification systems (one classifier per feature) trained using the first half TRECVid 2007 development data set. The fusion of classifiers outputs was finally provided by training a neural network based on evidence theory NNET [5] on the second half of the training data set.

Five runs are submitted using different types of descriptors provided by EURECOM, DCU, JRS and TUB:

1. Run 1: MPEG-7 global descriptors.
2. Run 2: MPEG-7 region descriptors.
3. Run 3: Combination of MPEG-7 global with TUB face detector and JRS motion activity descriptors.
4. Run 4: Combination between all descriptors (DCU MPEG-7 global and region, TUB face detector and JRS motion activity descriptors).
5. Run 5: Color and texture descriptors are extracted using three segmentation methods (A fixed image grid, watersheds [21] and a technique based on Minimum Spanning Trees MST [8]).

3.2 Institut TELECOM features

We have proposed the same audio features as last year [23]. These features are deduced after the outputs of an audio classification system which is designed to discriminate 17 different classes of sound, namely clean speech, noisy speech, music, music and speech, silence/pause and various environmental sounds (*i.e.* airplane, helicopter, applause, crowds, dogs, explosion, gun-shot, car, race-car, siren, truck/lorry/bus, motorcycle). The fraction of each class positive outputs over a video shot length are used as audio features.

3.3 ITI

For the detection of 3 other high-level features, namely *Building*, *Car* and *Waterscape-Waterfront*, a *Support Vector Machines* (SVMs) structure made of 3 individual SVMs was utilized, exploiting four MPEG-7 descriptors. The common TRECVID annotations were employed for their training. At the evaluation stage, for every keyframe the extracted low-level descriptors were combined and their values were normalized; these served as input to each SVM, which at the evaluation stage returned for every keyframe a numerical value in the range $[0, 1]$. This value denotes the degree of confidence with which the corresponding keyframe is assigned to the high-level feature associated with the particular SVM. For its calculation, the distance z from the corresponding SVM's separating hypersphere was estimated and subsequently mapped to $[0, 1]$ with the use of a sigmoid function. More details on the approach followed can be found in [23].

3.4 NTUA

In this section we describe the approach [20] followed for the detection of 10 high-level features, namely *Desert*, *Road*,*Sky*, *Snow*, *Vegetation*, *Explosion/Fire* and *Mountain*. The first step was to extract the following MPEG-7 descriptors, which from all the available NRKF keyframes: *Scalable Color*, *Homogeneous Texture*, *Edge Histogram* and *Color Layout*. Then, all images are segmented in a way that coarse segments are produced. The aforementioned MPEG-7 descriptors are then extracted from each image region. K-means clustering is performed on the descriptions of all regions of the training set, with the number of K set to 100. After this process, each cluster may or may not represent a high-level feature and each high-level feature may be represented by one or more clusters.

From each one of the formed clusters, the region that lies closest to the centroid is selected and will be referred to as "Region Type". An image will then be described semantically in terms of the region types it is consisted of. Next, for each one of the keyframes, a model vector is formed. More specifically, let: $d_i^1, d_i^2, ..., d_i^j, i = 1, 2, \dots, N_R$ and $j = N_C$, where N_C denotes the number of region types, N_R the number of the regions within the image and d_i^j is the distance of the i-th region of the image to the j-th region type. The model vector D_m is formed in the way depicted in equation 1.

$$D_m = \left[min\{d_i^1\}, min\{d_i^2\}, ..., min\{d_i^{N_C}\}\right], i = 1, 2, \dots, N_R \qquad (1)$$

For each semantic concept, a separate neural network-based detector is trained. Its input is the model vector and the output represents the distance of each region to the corresponding semantic concept. For the training of these detectors the common annotation has been used.

3.5 JRS

For content analysis, JRS contributed with the extraction of a number of visual indexes, as described in our last year's paper [23]. For each shot, we extracted features for visual activity, shot and keyframe similarity and occurrences of faces. Further camera motion (pan, zoom and tilt) [3] was computed by trajectory clustering.

The description of all feature extraction results is in MPEG-7 format compliant to the Detailed Audiovisual Profile (DAVP) we have specified [2].

3.6 QMUL

For feature extraction, QMUL contributed with the extraction of the following three features: "Sky", "Weather" and "Maps". The features are extracted using Biologically inspired classifier namely Particle Swarm Optimization (PSO). PSO is one of the evolutionary computation techniques. It was originally inspired by the social behavior of a flock of birds [12]. The image classification is performed using the Self Organizing Maps (SOM) and optimizing the weight of the neurons by PSO. A overview of the technique used is presented in [7] and was also used in last year's activity [23].

3.7 TUB

The feature extraction methods developed for TRECVID 2007 submission [23] were applied to the TRECVID 2008 data. Audio classification/segmentation and speaker change detection have been applied to the audio data. For the visual data faces and text regions have been detected for each keyframe. The goal of the *audio classification/segmentation* module is to split the audio stream into temporal segments and classify each segment into 6 predefined audio classes (pause, clean speech, noisy speech, pure music, music and speech, environmental sound). Therefore the audio stream is divided into nonoverlapping frames which are described by Mel frequency cepstrum coefficients (MFCC). For each of the classes a Gaussian mixture model (GMM) is trained. These models are used together with the maximum likelihood (ML) decision rule to classify the audio frames into the different categories. The goal of the *speaker change detection* is detect change points between individual speaker within the speech segments of the audio stream. This is achieved by dividing the audio stream into nonoverlapping frames which are described again by MFCCs and detecting change points by applying the Bayesian information criterion (BIC). The goal of the *face detection* module is to detect and localize frontal faces within the representative (RKF) and non representative keyframes (NRKF) of a shot and extract face statistics (number of faces, size of the largest face). Therefore the holistic face detection approach by Viola & Jones [22] was adopted, which is based on a combination of Haar features and a Adaboost trained classifier cascade. The *text detection* module detects and localizes text regions within the non representative keyframes of a shot and provides their number and location forsubsequent analysis. It is based on the joint analysis of edges and motion within the shot to detect static text regions.

3.8 UEP - Keyframe similarity using LSI with enhanced sparse image representation

The Latent Semantic Indexing (LSI) method [9] was developed for the automated information retrieval of large amount of text documents especially because of efficient matching of polysemy and synonymy. We extended the original LSI for intelligent image retrieval [16]. Our previous approach produced large non-sparse document matrices, because a raster image was coded as a sequence of pixels [18].

Although images can be represented sparsely for instance by the Discrete Cosine Transform (DCT) coefficients, the sparsity character is destroyed during the LSI-based dimension reduction process. In our approach, we keep the memory limit of the decomposed data by a statistical model of the sparse data. The aim is to find a small but "important" sub-set of coefficients, which represent semantics of images efficiently. The description of the algorithm is presented in [17]. In TrecVid 2008, we represented each keyframe by a vector with only 51 dimensions, see Table 1.

Properties of the document matrix *A*	
Number of keywords:	51
Number of documents:	1 422
Size in memory:	0.084 MB
The SVD-Free LSI processing parameters	
Dim. of the original space	1 422
Dim. of the reduced space (k)	15
The total time	0.28 secs.

Table 1: An example of image retrieval using the SVD-free Latent Semantic Indexing method related; Properties of the document matrix (up) and LSI processing parameters (down).

Having increased the computational effort of the LSI-based dimension reduction technique did not imply the increasing quality of retrieved results [17]. For this reason, only 15 eigenvalues and corresponding eigenvectors of the sparse partial symmetric eigenproblem was finally computed and stored in memory. This make here presented automated sparse image retrieval approach very efficient in sense of the computer memory and computer time.

3.9 DCU - Retrieval Engine

The common search engine leverage multiple modalities to form a response to an information need from a user. The search engine allows for multiple query by example, text queries and mixed modality queries. For visual components of queries we made use of six global visual features identified earlier, ranking within each was handled by the similarity measures as specified by the MPEG7 specification [13]. These measures for the most part are similar to Euclidian distance.

The previous content-analysis techniques could be accessed via two mechanisms within the search engine. The first method was to use the outputs of the previous methods as 'filters' on a result set of shots. The filters could have three states, 'show only shots matching the filter', 'shot shots not matching the filter' and no effect (default).

The second method of access incorporated not only the K-Space content-analysis results, but also results from the CU-VIREO374 collection donated by City University of Hong Kong and Columbia University [11] for which we are very grateful. We took the names of the concepts detectors and ran these through wordnet obtaining the synonyms for these terms. Therefore for each shot we had a bag of words which described the visual aspects of that shot. This text for each shot was then augmented with the translated ASR text provided by the University of Twente [10]. This therefore produced for each shot a collection of terms which described the content of the shot incorporate both visual and audio information. The text was then indexed by Terrier [15], with retrieval results provided through a vector space model.

When specifying a multi-modal query, the user can select to use any or all of these seven experts to retrieve a response. When a query is issued, it goes to each of the retrieval experts and a ranked list is returned. Using a variation on DCU's query-time weight generation techniques [24], these result lists are merged at query time with weights being assigned to each expert which approximate that experts likelihood of providing the most relevant responses to the query.

4 Interactive Search

This section will detail our experimental design including topic arrangement and qualitative data gathering. We will follow this with a brief explanation of the three systems used in the experiment.

4.1 Experiment Design

Our experiment was conducted over three geographic sites, all research laboratories. Each site obtained 12 users to conduct the experiment. Every user was to complete 12 of the 24 topics performing 4 topics on each of the three systems. Assignment of topics a user to complete was as specified by a latin squares arrangement, shown in Figure 1. In this diagram, users are the rows, with topics in columns.

The topic progression for a user was that they completed all topics for any given system first before progressing to the next system. The order of the presentation of the interfaces was randomized. We did not however randomize the order of the topics for a given system. Therefore during the completion of topics for a given system a learning effect should be observable. By not randomizing the topic order within a system we are able to measure this across users and sites. For example, in Figure 1 user 'S1' is scheduled to complete topics 221-224 on system GU, topics 229-232 on system DCU-1 and finally topics 241-244 on system DCU-2. The order in which the user uses the systems is randomized, so user 'S1' may have completed DCU-2 first, then GU and DCU-1, however within those systems the order in which they completed the topics is fixed, e.g. for DCU-1 they would have completed first topic 229, then 230 etc.

For every user we extensively logged both qualitative and quantitative data. The qualitative data we gathered for every user was at the pre-experiment, post-system training and post-topic level, providing an extensive amount of feedback. Furthermore several institutions conducted informal interviews after the experiments with the users to solicit further feedback. For quantitative analysis we logged multiple activities of the user including events such as searches, shots saved, shots played and shots removed.

The users we obtained for these experiments were a mix of college level students and research lab personnel. All users would be classed as 'non-expert' as none were involved in the design or implementation of any of the search interfaces used. Furthermore our mix of users included individuals who would not have had prior exposure to multimedia content-based retrieval.

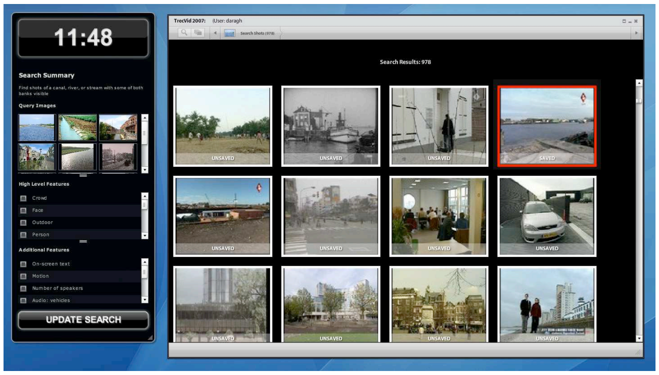
The following two subsections provided a brief overview of the DCU interfaces developed. A more complete explanation can be found in [6].

4.2 Interface DCU-1, Shot based

The 'shot based' system presented to the user the ranked list of shots direct from the retrieval engine. Presented in Figure 2 it can be seen that the ranked shots are organized left to right, top to bottom. It can be thought of as the more traditional result display that has been used for content-based retrieval interfaces. This interface displays no context for any of the returned results.

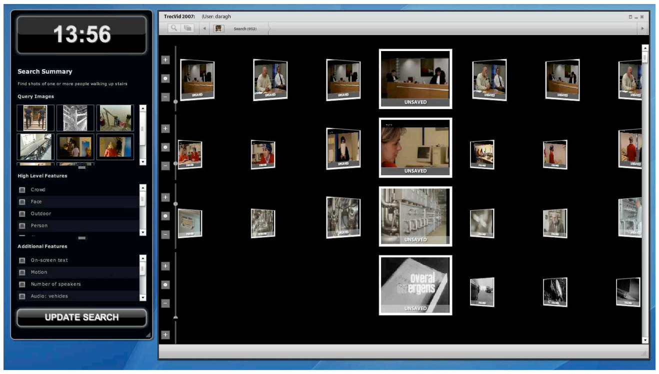
4.3 Interface DCU-2, Broadcast based

The 'broadcast based' system takes the idea of context to its maximum by ranking not shots, but broadcasts. If we were to take the assumption that the corpus for this year will have broadcasts which are more homogeneous on a subject (i.e. a documentary may be about one major subject, whilst in previous years a news broadcast could be seen as containing many subjects), then ranking broadcasts as opposed to shots appears as an interesting alternative. In Figure 3 we can see a horizontal line of shots in rows across the results area. Each of these rows is a ranked entire broadcast, with the best-matching broadcast being the first row. When a user issues a query, the ranked list of broadcasts is presented, and within each broadcast's row the row will be centered on the highest matching shot within that broadcast.



Find shots of a canal, river, or stream with some of both banks visible

Figure 2: Shot-based user interface



Find shots of one or more people walking up stairs

Figure 3: Broadcast-based user interface

4.4 Interface GU, Zooming Interface

Based on experience gained when carrying out different user studies with previous video retrieval interfaces, we found that users spent a considerable length of time browsing videos to look for relevant material. The value and importance of a search result appears to be based on it's value as a good starting point for a user to find other relevant shots within a video, by browsing the video, as much as the relevance of the result itself. Based on the ability of users to easily browse videos, and the willingness of many users to do so to find relevant material, we are proposing an interface that (a) emphasizes results which are good starting points from which to find material (point-finding within videos), and (b) extend the video browsing elements of the user interface to enable users to more easily view and browse videos. In order to achieve point (a) above we introduce a diversity based reranking into our search results to present more starting off points for browsing, and in order to achieve point (b) we introduce a zooming interface to help users explore and view more of a videos content in order to extend the range of material viewed by users when engaged in neighborhood search.

4.4.1 Diversity

Our diversity measure uses low level features in order to re-rank results, the strategy incrementally building a new and more diverse set of results from an initial result set. During each step keyframes are re-ranked according to their "quality", with the highest "quality" keyframe being added to the new result list. A Greedy selection algorithm [19] uses a quality metric (see Equation 2) which combines the similarity between the query t and each keyframe in the results set, with the dissimilarity between the keyframe in the result set with the re-ranked result set (see Equation 3). The first keyframe in the re-ranked result list is always the same as the original result list. For each iteration that follows the keyframe selected is the one with the highest combination of similarity to the original query and diversity relative to the re-ranked result list.

$$Quality(t, z, R) = Similarity(t, z) * RelDiversity(z, R) \tag{2}$$

$$RelDiversity(z, R) = 0 \text{ if } R = \{\} \tag{3}$$

$$= \frac{(\sum_{i=1}^{m}(1 - Similarity(z, r_i)))}{m}, \text{ otherwise}$$

4.4.2 Zooming

The layout of our zooming interface is inspired by the map metaphor, used by websites such as Google Maps[1] and Multimap[2]. The map metaphor is useful, since in such interfaces it is possible to zoom in and out to see more or less of the map at a lesser or greater scale, i.e. to trade detail for an overview of a larger area of ground. In a similar way, zooming functionality allows the user to see more or less of the neighborhood of shots in varying degrees of detail. Additionally, a grid organization enables the interface to display more of the neighborhood of shots around a central starting point. We ignore other issues at the moment, such as the display of automatic speech recognition text and playback of videos, in order to focus on range extension. By default we displayed a 5 by 5 grid of keyframes, with the temporal order of the keyframes starting at the top left of the display, and ending at the bottom right. When a keyframe is clicked the display will centre on that keyframe and update to show the neighboring keyframes. This allows a user to browse forward and backwards in a video by up to 12 keyframes in either direction. The interface in Figure 4 zooms by increasing the degree of granularity represented by each keyframe. The bar at the bottom of the display indicates the current zooming level, the plus and minus buttons on either side allows the user to zoom in or out. The mouse scroll wheel can also be used to adjust the zoom level. When the user zooms out one level, each keyframe will represent two shots, if they zoom out again, each keyframe will then represent 4 shots, then 8 shots, and so on as the user zooms further and further out. In effect, each zooming out will remove half of the keyframes on the display, while bringing into view other shots from further away in the video.

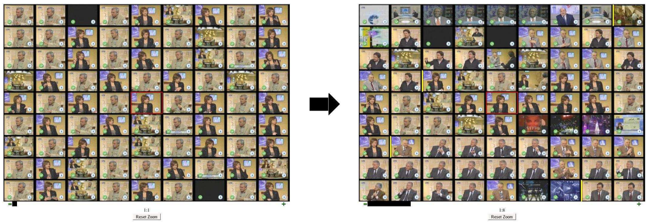


Figure 4: Zooming Interface, unzoomed interface on the left, and examples of zooming out on the right.

5 Results

In this section we will present some preliminary results and analysis of our results from the experiment for 2008. As we produced 18 runs in total with three interfaces over three sites the amount of data we have produced leads us to relatively cautious analysis at this very early stage. Officially to NIST we submitted three runs. The creation of these three runs was for each system to take the outputs of the user who saved the most for a given topic. The analysis we are conducting here is primarily centered around the number of shots saved by a user for a given topic on a given system. We do not take into account in this analysis if the shots being saved were judged as relevant by NIST as we make the assumption for this initial analysis that a user believed the shot to be relevant for the stated information need.

Our initial analysis is presented in Figure 5. In this diagram there are four graphs. The bottom right graph is a plot of InfAP of the official runs across topics. The remaining three diagrams display for every user how many shots they saved when compared against the mean number of shots saved for that system, normalized as standard scores which gives how far away from the mean in terms of standard deviations the user was. There is one graph for each of the interfaces developed. Taking the graph for DCU-1, if we examine topic '0221' we can see that of the six users for that topic (2 each from CWI, DCU and GU), that one CWI user and the two DCU users saved more than the average number of shots for that system for that topic, whilst one CWI user and the

[1] http://maps.google.com
[2] http://www.multimap.com

two GU users saved less than the average number of saved shots for that topic.

As much of this analysis is still very much at a preliminary stage we are leaving much of the interpretation of these results to the reader as multiple conclusions can be drawn. Initial observations that can be made is that the DCU users on average saved more shots than the GU users. Furthermore, as the presentation of these results is Z-Scores, which is displaying how many shots were saved in terms of standard deviations, we note that the Glasgow system appears to be more compact in terms of the range of the standard deviations. This indicates that users of the Glasgow system typically saved similar numbers of shots for a given topic. Conversely the system DCU-2 (broadcast based system) exhibited a greater variance in user performance of saved shots.

Figure 6 is an initial aggregation of results to determine if any large effects can be observed. We present here two graphs. The first graph titled 'Std Dev by Site' is an attempt to examine the effect individual sites can bring to a search experiment. For this graph we have aggregated the number of shots saved for every user for every system at each indiviual site, then calculated from those users what the standard deviation with respect to number of shots saved was for every topic. For example, for DCU, we obtained the standard deviation of the number of saved shots for topic '0221' calculating this across the six users who did the first topic (two users on each of the three systems). We repeated this for CWI users and GU users. This gave us for each topic for each site a single figure which defined how much variability there was amongst users at that site, ignoring the effect of different interfaces. For each of the three sites we then determined the releative difference between them and that is what is presented in the graph. Taking topic '0221' we see a value of '0.43' for CWI, '0.53' for DCU and '0.04' for GU. This means that for topic '0221' DCU displayed the greatest variance in saved shots amongst users releative to CWI and GU. Another trivial example would be that if for a given topic all sites had values of '0.33' it would mean that for that topic each of the sites exhibited the same variance in number of shots saved from its users as the other two sites. From an initial inspection of this graph, we can see that for a majority of topics sites performed similary to each other with respect to variance in its users of number of shots saved. This is a positive indicator that differences in the interfaces are producing similar differences across sites. Again however we would note the very preliminary nature of these observations.

The second graph (titled 'Std Dev by System') shown is similar to the previous graph in that it is again an aggregation which is showing the variance in the number of shots saved. However in this graph the aggregation is by system, rather than by site. For example, for topic '0221' we calculated the standard deviation of shots saved by users of system 'DCU-1' across sites, and repeated this for 'DCU-2' and 'GU' systems. Then we calculated the releative differences in the standard deviations from these systems. In this graph, system 'DCU-1' exhibits the greatest variance in number of shots saved by its users, with DCU-2 having the lowest. There are multiple observations that can be made from this data. The first is that as the users completed the topics in blocks of 4 in a fixed order (as described earlier) we can observe that generally there is an increase in the variance of shots saved as users learned how to use the system they were using at the time. This presents itself as a gradual increase in variance over blocks of 4 topics (e.g. topics 0221-0224, 0237-0240 etc). This is encouraging as it appears to be evidence of a learning effect in action and will allow deeper exploration of which systems produced a greater learning effect in its users and which systems the user exploited better initially in achieving better performance. The second observation we can make is that generally it appears that system 'DCU-2' (broadcast based system) produces the greatest variance in its number of saved shots than the other two systems. This is interesting as it would appear to correlate with the graph for 'DCU-2' in Figure 5 which showed that users exhibited the largest range of variance in number of shots saved.

As stated previously this is only an initial analysis of the results we have obtained. Future work in examining this data set will examine multiple aspects of the search experience from a user perspective. Given the large range of aspects that could possibly be examined it is our intention to release this data set to the wider multimedia retrieval community in the near future.

6 Conclusion

We have presented the K-Space participation in TRECVid 2008. This was our third and final participation in TRECVid. We undertook one of the largest laboratory video retrieval experiments to date and have presented the composition of the systems used for this activity. We have also conducted a preliminary analysis of the results we obtained from this activity. Given the nature of this data we believe it would be of benefit for this data to be made to the wider multimedia retrieval community and it is our intention to publicily release this data in the near future.

7 Acknowledgments

The research leading to this paper was supported by the European Commission under contract FP6-027026 (K-Space).

References

[1] KSpace Network of Excellence, information at http://www.k-space.eu/.

[2] W. Bailer and P. Schallauer. The Detailed Audiovisual Profile: Enabling Interoperability between

MPEG-7 based Systems. In *12^{th} International MultiMedia Modelling Conference (MMM'06)*, pages 217–224, Beijing, China, 2006.

[3] W. Bailer, P. Schallauer, and G. Thallinger. Joanneum Research at TRECVID 2005 – Camera Motion Detection. In *Proceedings of TRECVID Workshop*, pages 182–189, Gaithersburg, Md., USA, 11 2005. NIST.

[4] R. Benmokhtar, E. Galmar, and B. Huet. Eurecom at TRECVid 2007: Extraction of high level features. In *TRECVid'07, 11th International Workshop on Video Retrieval Evaluation, Gaithersburg, USA*, November 2007.

[5] R. Benmokhtar and B. Huet. Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content. In *MMM, International MultiMedia Modeling Conference, January 9-12 2007, Singapore - Also published as LNCS Volume 4351*, January 2007.

[6] D. Byrne, P. Wilkins, G. Jones, A. F. Smeaton, and N. O'Connor. Measuring the impact of temporal context on video retrieval. In *CIVR 2008 - ACM International Conference on Image and Video Retrieval*, 2008.

[7] K. Chandramouli. Particle swarm optimisation and self organising maps based image classifier. In *Second International Workshop on Semantic Media Adaptation and Personalization*, pages 225–228, December 2007.

[8] P. Felzenszwalb and D. Huttenlocher. Efficiently computing a good segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–104, 1998.

[9] D. Grossman and O.Frieder. *Information retrieval: Algorithms and heuristics.* Kluwer Academic Publishers, Second edition, 2000.

[10] M. Huijbregts, R. Ordelman, and F. de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of the second international conference on Semantics And digital Media Technologies (SAMT)*, Lecture Notes in Computer Science, Berlin, December 2007. Springer Verlag.

[11] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo. CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection. Technical report, Columbia University, August 2008.

[12] J. Kennedy and R. C. Eberhart. *Swarm Intelligene.* Morgan Kaufmann, San Mateo, CA, 2001.

[13] MPEG-7. Multimedia Content Description Interface. Standard No. ISO/IEC nï¿½15938, 2001.

[14] N. O'Connor, E. Cooke, H. le Borgne, M. Blighe, and T. Adamek. The AceToolbox: Low-Level Audiovisual Feature Extraction for Retrieval and Classification. In *2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, 2005.

[15] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in terrier. *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, 2007.

[16] P. Praks, J. Dvorský, and V. Snášel. Latent semantic indexing for image retrieval systems. In *SIAM Linear Algebra Proceedings, Philadelphia, USA*. International Linear Algebra Society (ILAS), http://www.siam.org/meetings/la03/proceedings/-Dvorsky.pdf, July 2003.

[17] P. Praks, E. Izquierdo, and R. Kučera. The sparse image representation for automated image retrieval. Work in progress, 2008.

[18] P. Praks, L. Machala, and V. Snášel. *On SVD-free Latent Semantic Indexing for Iris Recognition of Large Databases.* Springer-Verlag London, In: V. A. Petrushin and L. Khan (Eds.) Multimedia Data mining and Knowledge Discovery (Part V, Chapter 24), 2007.

[19] B. Smyth and P. McClave. Similarity vs. diversity. In *ICCBR '01: Proceedings of the 4th International Conference on Case-Based Reasoning*, pages 347–361, London, UK, 2001. Springer-Verlag.

[20] E. Spyrou, G. Tolias, P. Mylonas, and Y. Avrithis. A semantic multimedia analysis approach utilizing a region thesaurus and LSA. 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008), 2008.

[21] L. Vincent and P. Soille. Watersheds om digital spaces: An efficient algorithm based on immersion simulations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13(6), June 1991.

[22] P. A. Viola and M. J. Jones. Robust real-time object detection. In *IEEE Workshop on Statistical and Computational Theories of Computer Vision*, 2001.

[23] P. Wilkins and et al. KSpace at TRECVid 2007. In *TRECVID 2007 – Text REtrieval Conference, TRECVID Workshop, Gaithersburg, Md., 5-6 November 2007*, 2007.

[24] P. Wilkins, P. Ferguson, and A. F. Smeaton. Using score distributions for querytime fusion in multimedia retrieval. In *MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
S1	GU	GU	GU	GU					DCU_1	DCU_1	DCU_1	DCU_1									DCU_2	DCU_2	DCU_2	DCU_2
S2	DCU_1	DCU_1	DCU_1	DCU_1					DCU_2	DCU_2	DCU_2	DCU_2									GU	GU	GU	GU
S3	DCU_2	DCU_2	DCU_2	DCU_2					GU	GU	GU	GU									DCU_1	DCU_1	DCU_1	DCU_1
S4	GU	GU	GU	GU									DCU_1	DCU_1	DCU_1	DCU_1	DCU_2	DCU_2	DCU_2	DCU_2				
S5	DCU_1	DCU_1	DCU_1	DCU_1									DCU_2	DCU_2	DCU_2	DCU_2	GU	GU	GU	GU				
S6	DCU_2	DCU_2	DCU_2	DCU_2									GU	GU	GU	GU	DCU_1	DCU_1	DCU_1	DCU_1				
S7					GU	GU	GU	GU	DCU_1	DCU_1	DCU_1	DCU_1					DCU_2	DCU_2	DCU_2	DCU_2				
S8					DCU_1	DCU_1	DCU_1	DCU_1	DCU_2	DCU_2	DCU_2	DCU_2					GU	GU	GU	GU				
S9					DCU_2	DCU_2	DCU_2	DCU_2	GU	GU	GU	GU					DCU_1	DCU_1	DCU_1	DCU_1				
S10					GU	GU	GU	GU					DCU_1	DCU_1	DCU_1	DCU_1					DCU_2	DCU_2	DCU_2	DCU_2
S11					DCU_1	DCU_1	DCU_1	DCU_1					DCU_2	DCU_2	DCU_2	DCU_2					GU	GU	GU	GU
S12					DCU_2	DCU_2	DCU_2	DCU_2					GU	GU	GU	GU					DCU_1	DCU_1	DCU_1	DCU_1

Figure 1: Latin Square arrangement used for 3 systems

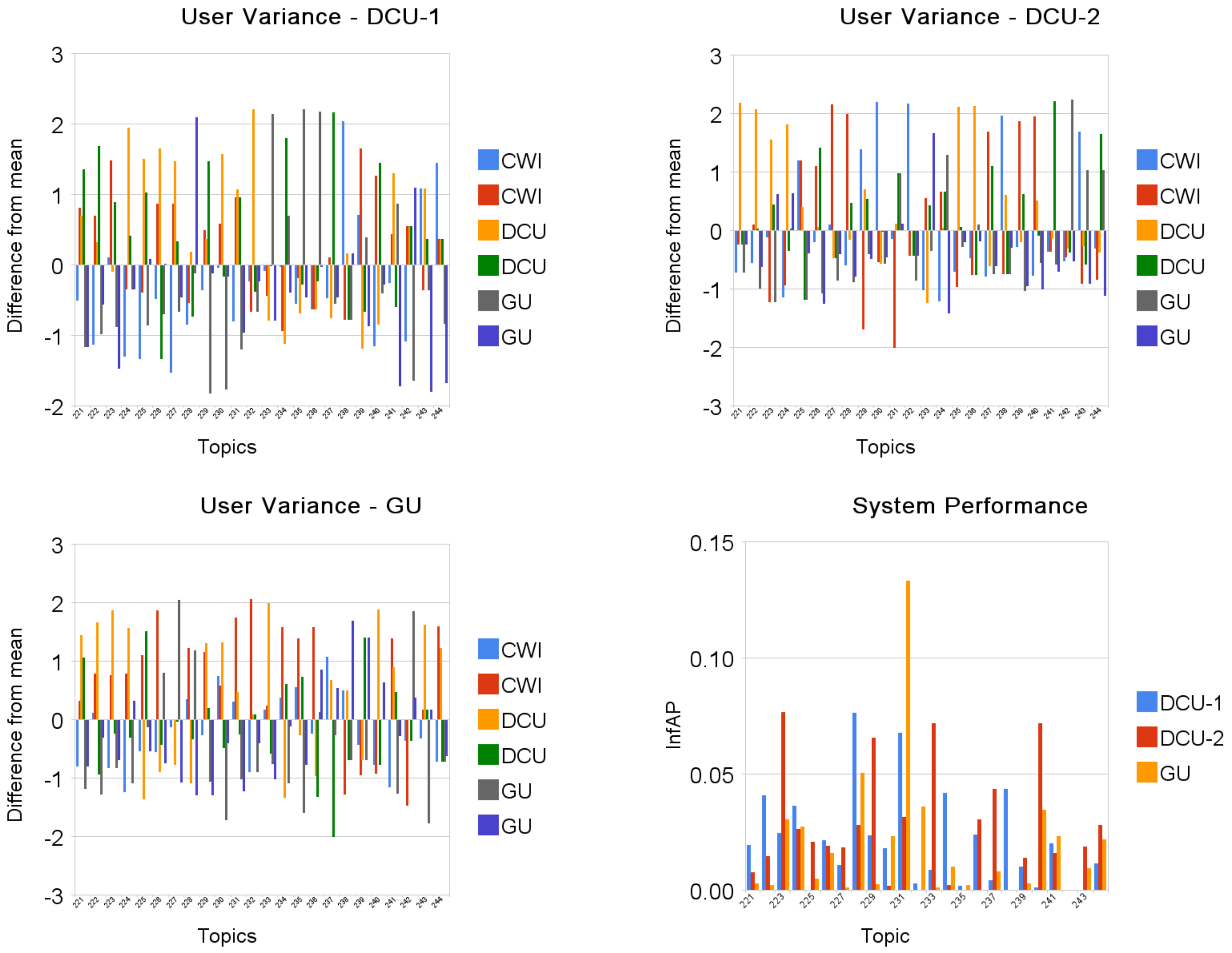


Figure 5: User Variance Across Sites and Official Performance

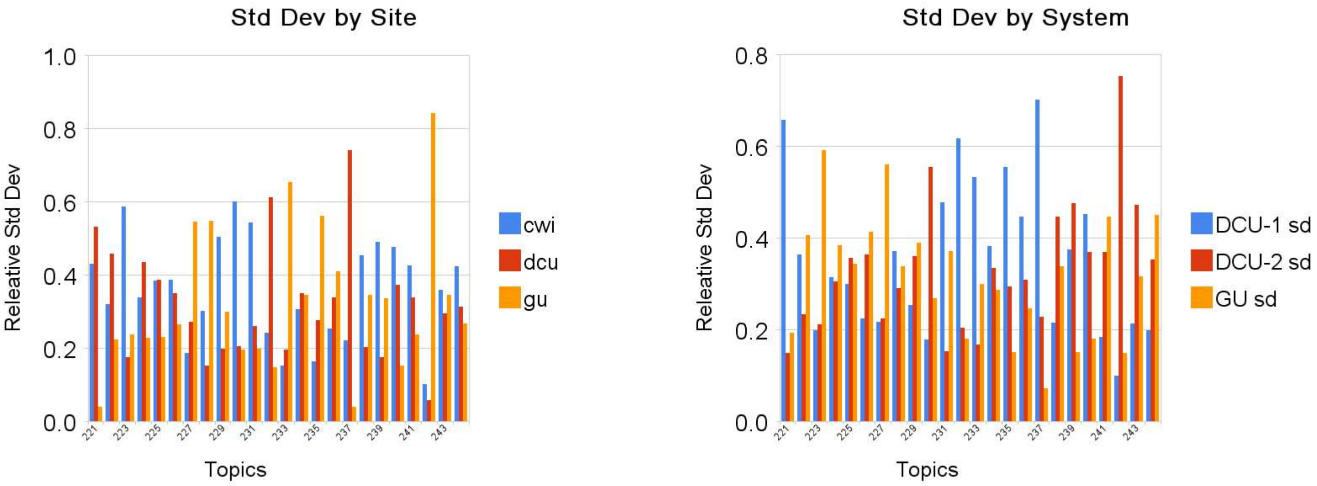


Figure 6: Standard Deviation by Site and System