AN EXTENSIBLE MODULAR COMMON REFERENCE SYSTEM FOR CONTENT-BASED INFORMATION RETRIEVAL: THE SCHEMA REFERENCE SYSTEM

Vasileios Mezaris[1,2]*, Haralambos Doulaverakis*[2]*, Stephan Herrmann*[3]*, Bart Lehane*[4]*,*
Noel O'Connor[4]*, Ioannis Kompatsiaris*[2]*, Walter Stechele*[3]*, and Michael G. Strintzis*[1,2]

[1]Information Processing Laboratory, Electrical and Computer Engineering Department,
Aristotle University of Thessaloniki, Greece
[2]Informatics and Telematics Institute/Centre for Research and Technology Hellas, Thessaloniki
[3]Institute for Integrated Systems, Munich University of Technology, Munich D-80290, Germany
[4]Centre for Digital Video Processing, Dublin City University, Ireland

ABSTRACT

The SCHEMA Reference System is a content-based image and video indexing and retrieval system that adopts a module-based, expandable architecture. Using this module-based approach, five different analysis modules, developed at different researcher laboratories, have been integrated with it. Within the system, the MPEG-7 XM (MPEG-7 eXperimentation Model), a non-normative part of the MPEG-7 standard realizing the normative descriptors, is employed along with extensions for descriptor extraction and for supporting search and retrieval functionalities. In addition to the XM, the system supports high level descriptors (e.g. face/non-face image categorization) and content-based indexing and retrieval using other modalities (e.g. pre-existing keyword annotations, text generated via automatic speech recognition (ASR)). In this paper, the TRECVID 2004 test corpus is used as a common data set for demonstrating the functionalities and the efficiency of the proposed system.

1. INTRODUCTION

Huge amounts of digital multimedia content are currently available in unstructured and non-indexed forms on the web, in proprietary commercial repositories (e.g. content houses, broadcaster archives) and not least personal collections (e.g. home movies, personal photographs). This explosion of content is the result of recent advances in a number of key enabling technologies.

The sheer size of content databases makes their efficient management an extremely challenging task. One key enabling functionality is the ability to index the content in terms of the semantics it represents and to provide access in a way that comes naturally to users. Typically, given the volume of information to be dealt with, it is desirable that this indexing take place in a completely automatic manner, or a least with a minimum of user interaction.

The goal of the SCHEMA Network of Excellence is to develop a software instantiation of a prototypical retrieval system. This paper focuses on a description of the design of the SCHEMA Reference System and an illustration of it's use as part of the US NIST TRECVID (Text Retrieval Conference – Video Track) initiative to benchmark information retrieval systems [1]. The overall system architecture is described in Section 2. The individual Reference System modules integrated thus far are described in Section 3, including contributed spatial segmentation modules (Section 3.1), modifications to the MPEG-7 XM (Section 3.2), high-level feature extraction modules (Section 3.3) and a textual information processing module (Section 3.4). The results obtained using an implementation of the Reference System developed for SCHEMA collaborative participation in TRECVID 2004 are presented in Section 4. Finally, some conclusions and directions for future work on the system are presented in Section 5.

2. REFERENCE SYSTEM ARCHITECTURE

The architecture of the SCHEMA Reference System is module based and inherently expandable. Clearly defined interfaces between different modules allow many different researchers to easily integrate contributed or proprietary modules. The design takes into account a formal study of the user and system requirements, including aspects such as response times, standardization and scalability of content-based information retrieval systems, that was carried out by the Network.

The system combines five analysis modules developed by different SCHEMA partners and affiliated members. In combination with the low-level descriptors extracted using

This work was supported by the EU project SCHEMA "Network of Excellence in Content-Based Semantic Scene Analysis and Information Retrieval" (IST-2001-32795).

ISBN: 952-15-1364-0

the output of segmentation, the system can also support high level (semantic) descriptors and the integration of content-based indexing and retrieval with other modalities (i.e. pre-existing keyword annotations, text generated via automatic speech recognition (ASR)). More specifically, it employs a high-level semantic classification algorithm for categorizing images into face and non-face classes (whereby each class indicates whether or not the image contains one or more human faces), a module for motion characterization, as well as a module for exploiting any available textual annotations or transcripts. It must be noted that the aforementioned modules are just examples of what can be integrated with the system; additional modules (e.g. sound analysis) could be integrated, depending on the application.

The MPEG-7 standard was adopted to allow standardized representation of the multimedia descriptors extracted by the analysis modules. The MPEG-7 standard [2], formally named ”Multimedia Content Description Interface”, provides a rich set of standardized tools to describe multimedia content. The proposed system uses the MPEG-7 XM [3] (MPEG-7 eXperimentation Model, a non-normative part of the Standard realizing the normative descriptors) for descriptor extraction and for supporting search and retrieval functionalities. It also employs several extensions to the MPEG-7 XM, which improve its efficiency and considerably extend its functionalities.

All the aforementioned functionalities have been combined under a common Graphical User Interface, built using web technologies. The resulting system constitutes an effective experimental platform for the evaluation and comparison of different analysis, indexing and retrieval modules. This reference design is publicly available [1] and allows other researchers the possibility to use the system as either a means of benchmarking their own systems or as an integration platform for their own indexing and retrieval modules. An overview of the proposed architecture is illustrated in Figure 1.

3. REFERENCE SYSTEM MODULES

In this section, a description of the different modules comprising the SCHEMA Reference System is presented. The presentation starts with the image segmentation algorithms that have been integrated. Following that, the process of indexing and retrieval using the MPEG-7 XM is discussed and a number of extensions to the XM, developed by SCHEMA to improve its efficiency, are illustrated. A description of the modules supporting high-level queries that have been integrated thus far, namely the high-level face/non-face classifier, the high-level motion features and the textual information processing algorithm, is also provided.

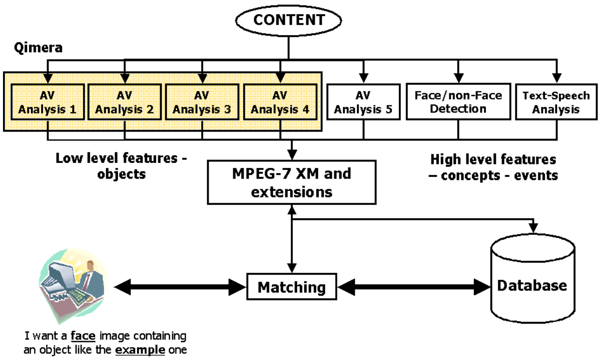


Fig. 1. Overview of the SCHEMA Reference System

3.1. Visual Content Analysis

A total of five image segmentation modules contributed by four SCHEMA partners and one affiliated member have been integrated with the Reference System. Four of these modules were previously integrated within the Qimera framework [4], which provides common input/output formats, thus facilitating the rapid subsequent integration of different segmentation modules with the Reference System; for the fifth one, the same input/output formats were used for integration with the Reference System. The segmentation modules integrated with the Reference System are:

- a Pseudo Flat Zone Loop algorithm (PFZL);
- a Modified Recursive Shortest Spanning Tree algorithm (MRSST);
- a K-Means-with-Connectivity-Constraint algorithm (KMCC);
- an Expectation Maximization algorithm (EM) in a 6D colour/texture space.
- a Watershed Segmentation and Rag Minimax algorithm (WSRM).

The system can use any of the above modules to produce region-based segmentations of images/key-frames prior to region-based indexing. The segmentations produced, are post-processed in order to eliminate any small undesirable regions and to restrain the maximum number of generated regions to 10, so as to avoid performing indexing and retrieval in an unnecessarily large region collection. Post-processing was based on merging undesirable regions in an agglomerative manner [5]. Illustrative results of the different segmentation modules are presented in Figure 2. Details can be found in [6], [7], [8] and [9].

[1] Schema Reference System. http://media.iti.gr/SchemaRS

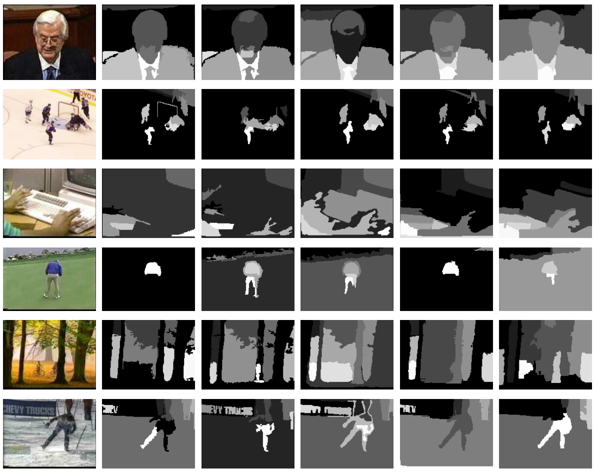


Fig. 2. Sample results of the integrated segmentation modules. The results generated by MRSST, KMCC, EM, PFZL and WSRM are presented in columns 2 to 6, respectively.

3.2. Indexing and Retrieval using the MPEG-7 XM and its Extensions

The MPEG-7 XM supports two main functionalities:

- Extraction of a standardized Descriptor (e.g. Dominant Color Descriptor) for a collection of images or image regions – this is termed the *extraction application*.

- Retrieval of images or image regions of a collection that are similar to a given example, using a standardized Descriptor and a corresponding matching function to evaluate similarity – this is termed the *search and retrieval application*.

While these functionalities are the basic building blocks of any content-based indexing and retrieval system, the direct use of their XM instantiation in a real-world system has certain drawbacks. These include the inability of both the extraction application and the search and retrieval application to consider more than one descriptor simultaneously and the reduced time-efficiency of retrieval. The latter is a result of the need to decode the binary encoded descriptions for the image collection during each query, as well as the lack of a true indexing mechanism. To address these drawbacks, extensions to the original XM software have been developed and are described in the sequel.

3.2.1. XM MultiImage module

The MultiImage module was developed to address the need to effectively combine more than one MPEG-7 descriptor. The MultiImage module implements both the extraction application, which extracts several of the MPEG-7 visual descriptors in order to generate a single .mp7 database file, and the MultiImage search application. The latter combines all the available descriptors to perform search and retrieval. To this end, default weights are defined for every descriptor used for the search. The MPEG-7 descriptors that are supported by the MultiImage module are *Color Layout*, *Edge Histogram*, *Color Structure*, *Homogeneous Texture*, *Dominant Color*, *Contour Shape*, *Scalable Color* and *Region Shape*.

Descriptors are instantiated using an abstract descriptor, the so-called MultiDescriptor for images. This abstract descriptor module simply encapsulates the memory management for the selected descriptors and allows calling the extraction and the matching of all the descriptors as if they were only a single descriptor. Figure 3 shows the architecture of the matching process using the abstract descriptor (marked in blue). When instantiating MultiDescriptor objects, the corresponding objects of the selected descriptors are also instantiated. This behavior is indicated by the grey (rounded) arrows. In a similar manner, when creating the overall processing chain, individual descriptor modules are also connected to their processing chains (marked with black arrows).

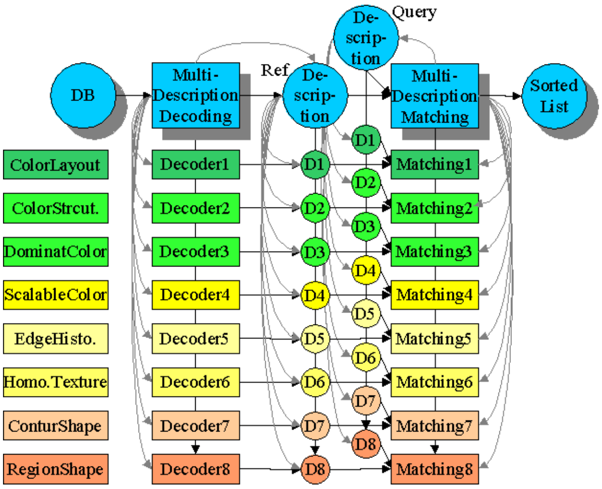


Fig. 3. Architecture of the MultiDescriptor module, i.e. the matching chain using multiple descriptors.

An essential problem when combining similarity values of different visual descriptors is the fact that the distance functions are not normalized in any way. Therefore, in these experiments a simple normalization of the working point of the matching functions was employed. The working point is defined as the distance value, below which descriptions can be treated as similar; if a distance value exceeds this threshold, the corresponding descriptions are assumed to be different. The threshold values for the individual descriptors were determined experimentally. Then, the distance functions can be simply scaled in a linear way, so that each working point has a distance value of 1.0. Then a simple weighted linear combination of the individual distance values is used to combine these normalized values. To date, these weights are not subsequently adjusted in any way; thus, all weighting factors have a value of 1.0. These could however be adjusted in the future using relevance feedback techniques.

3.2.2. XM Server

The original MPEG-7 reference software (XM software) is a simple command line program. When executing a similarity search using the selected visual descriptors, the program reads in the descriptions from the MPEG-7 descriptor bit stream. Then the query image is loaded and the query descriptions are extracted. Finally, the query description is compared to all descriptions in the reference database and the most similar descriptions are stored in a sorted list. The sorted list holds the n best matches only in order to simplify the sorting process. Using this command line tool means that for every search process the descriptions database is read and the query description is extracted. This leads to significant overheads in the search process, making a single search step slow. To accelerate the search procedure an extension to the original XM, termed *XM Server* was designed featuring the following modifications:

- First, it is not required to extract the query description from the image data if the query is already part of the description database. This kind of search is called ”intra search” and taking advantage of it can save a significant amount of overhead.

- A second modification is to keep the XM running in the background, accepting new queries and delivering the corresponding retrieval results. Thus, the process can be seen as a constantly running server application.

3.2.3. XM Indexing module

Although the implementation of the XM Server already leads to a significant speed-up of the search process, a linear search is still performed; in a huge database, the search could still be too slow. Profiles on a 1,5 GHz Pentium 4 showed that it is possible to perform 2000 matches of a descriptor set per second. This means, a search in a database of approximately 1 million images will take more then 8 minutes, making an interactive search useless in this case. A solution to speed-up the search process is to select a meaningful sub-set of the database and to perform the search only on this part. The selection of the right sub-set is done using an index structure.

This kind of indexing is an indexing of the descriptor data in contrast to an indexing of the media data (which is the descriptor extraction process). The index stores in the static index table pre-computed distance values between descriptors. Thus, re-using these distance values can help to accelerate the search procedure. The basic idea of the index is to create clusters of images that are similar to each other. In other words, the distance between the descriptors within a cluster is small, while the distances between descriptors of different clusters is large. The result is that during retrieval, only a subset of the images is compared which leads to greatly reduced search times.

In practical experiments using a database with nearly 800.000 elements three or four index levels were created. In this case, the time to perform the indexing was about 1 day (3 index levels, 1,5 GHz Pentium 4). When performing a search, typically a few thousand descriptors are matched. The number of matches can vary, because the index tree is not balanced with respect to the number of elements in the branches. Thus, the index reflects the bias of the database content. If there are many images showing mountains and sky, there will also be an accumulation in the index. On the other hand, the index is balanced with respect to the distances of the descriptor. This property enables to create descriptor databases with a constant density. These databases can be useful to normalize descriptor distance functions and for other applications.

3.3. High-level Features

3.3.1. Motion characterization

Two motion features are integrated into the SCHEMA Reference System, as examples of high-level features. The first is the MPEG-7 Motion Activity descriptor. This is a high-level descriptor, defined in MPEG-7 as being the standard deviation σ of the motion vectors in a video shot. MPEG-7 defines a five-notch qualitative scale for characterizing motion activity, ranging from “Very Low” ($\sigma < 3.9$) to “Very High” ($\sigma > 32$). More details on this can be found in [10].

The second motion feature is a measure of how much camera movement (pans, zooms etc.) is contained in each shot. The technique for estimating this feature examines the amount of consecutive zero motion vectors in each MPEG P-Frame in a shot. For a frame with camera movement, there should be very few of these zero-value runs, as most motion vectors will have a large value (i.e. contain movement). However, a frame with no camera movement will contain a large number of these runs. A threshold is used to determine if a frame can be declared as being a frame with camera motion, or a static frame. Finally, the percentage of frames with camera motion for each shot is found and this value serves as a measure of global camera movement. The correspondence between the numerical values of this feature and the values of a corresponding three-notch qualitative scale (”High”, ”Medium”, ”Low”) used in the SCHEMA Reference System is estimated using a Fuzzy C-Means algorithm.

3.3.2. Face/non-face image classification

The architecture that was adopted in terms of the implementation and testing of a high-level face/non-face classifying system is based on the model proposed in [11], [12]. However, the insertion of an additional step in the process of classification is proposed. Instead of applying the classification algorithm on the images, we first apply an automatic image segmentation algorithm (i.e. one of the segmentation algorithms already integrated with the SCHEMA Reference System) and then classify the resulting regions. Those regions are homogeneous in terms of color and texture, so they tend to correspond to meaningful entities.

More specifically, a set of classes $C = \{\omega_1, \omega_2, \ldots, \omega_N\}$ is initially defined, so that the classification problem becomes specific. In our case, the two defined classes $\omega_1,\ \omega_2$ represent face regions and non-face regions respectively. A set E containing both face images and non-face images is formulated, in order to be used as training and test set. The automatic image segmentation algorithm is then applied on that set of images, thus a set of regions P is produced. A number of standardized MPEG-7 low-level descriptors are then extracted, in order to serve as classification features for each region $p \in P$. The features that are used by the employed high-level classifier are the Dominant Color, Edge Histogram and Contour Shape descriptors; subsets of these features are incorporated into feature vector x, used for classification.

A set $P_{tr} \in P$ of regions which are manually classified (i.e. a semantic tag is attached to each one of them) is formed, in order to serve as a training set for the classifier, i.e to allow the estimation of the parameters of the multivariate Gaussian distribution used for modelling the probability $p(x|\omega_i)$. The initial set of regions P can subsequently be classified by use of a variation of the minimum error Bayes rule, the Neyman-Pearson rule

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \mu \rightarrow x \in \Omega_1 \qquad (1)$$

where the threshold μ is chosen empirically and Ω_1, Ω_2 are the regions in which the measurement space is divided by the decision rule; if an observer vector lies in region Ω_i, it is assumed to belong to class ω_i. Finally, a rule is used to classify the images based on the classification of the regions that constitute them

In order to test this classifier, a set of 616 images were used; these images can be roughly divided in three groups: a) images where the presence of human face is dominant (i.e. close-up face images), b) images where human faces are clearly distinguishable but not dominant, and c) images in which no faces are depicted. The results that were obtained are recorded in Table 1, while several sample results are depicted in Figure 4. These results indicate that the proposed region-based classification to face/non-face images is more accurate that the widely used global image classification. However, in the former case accurate classification may be affected by the accuracy of segmentation, as shown in Figure 4.

Tab. 1. Correct classification rates of face detection experiments

Region-level approach (using 2 Dominant Colors and Contour Shape, $\mu = 130$)		
Face images		Non-face images
$352/414$ (85%)		$146/202$ (72.3%)
Close-ups	Distant shots	
$184/214$ (85.9%)	$168/200$ (84%)	
Global image classification (using 2 Dominant Colors and Edge Histogram, $\mu = 5$)		
Face images		Non-face images
$290/414$ (70%)		$176/202$ (87.1%)
Close-ups	Distant shots	
$179/214$ (83.6%)	$111/200$ (55.5%)	

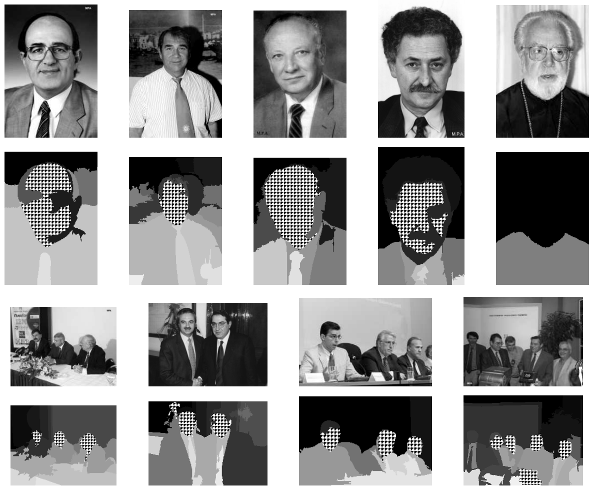


Fig. 4. Sample face detection results: the detected faces are the textured areas in the segmentation masks.

3.4. Textual Information Processing

The text ranking algorithm integrated with the SCHEMA Reference System is a parameterized one, incorporating both normalized document length (the associated text for every image/key-frame, in our case) and term frequency [13].

Let N be the number of documents in the entire collection, n be the number of documents query term i occurs in, $DL(j)$ be the length of document j that is to be scored, and ADL be the Average Document Length of all the documents in the collection. Then, $NDL(j)$ normalizes the length of document j by the Average Document Length and is defined as $NDL(j) = DL(j)/ADL$. The Collection Frequency weight for a query term i is defined as

$$CFW(i,j) = \log N - \log n \tag{2}$$

Denoting $TF(i,j)$ the Term Frequency of the i-th query term, i.e. the number of times term i occurs in document j, the overall score $S(i,j)$ of document j for query term i is calculated by

$$S(i,j) = \frac{CFW(i,j) \cdot TF(i,j) \cdot (K+1)}{K \cdot ((1-b) + b \cdot NDL(j)) + TF(i,j)}, \tag{3}$$

where K and b are application-specific constants. More specifically, K is used to modify the influence that the term frequency $TF(i,j)$ has on the final score. A value of $K = 0$ eliminates this influence, while larger values will result to $TF(i,j)$ having increasing impact. b, where $0 \leq b \leq 1$, modifies the influence of the document length on the score. For the TRECVID application of the SCHEMA Reference System, described in section 4, values $K = 2$ and $b = 0.75$ were used, since these are reported in [13] to produce good results.

4. EXPERIMENTAL RESULTS

The SCHEMA Reference System described in the previous sections was used for building a search and retrieval application, to perform indexing and retrieval in a dataset of approximately 33000 key-frames. These correspond to the 64 hours of news video used for the TRECVID 2004 experiments. The key-frames, each representing one video shot, are accompanied by transcripts of the spoken content of the corresponding shot, generated by means of Automatic Speech Recognition. The existence of textual information that is associated with the visual data, the possibility to extract motion information from the shots and the extensive test corpus containing both people-centric and non-people-centric segments make this an ideal test scenario for the combined use of all modules integrated thus far into the SCHEMA Reference System. The use of all 5 segmentation modules for segmenting each key-frame was adopted for the purpose of alleviating the imperfection of any segmentation algorithm.

In the specific application of the SCHEMA Reference System developed for this test corpus, a query can be performed in three ways (options):

- Using user-supplied keywords to evaluate shots with the help of the text-ranking algorithm (section 3.4) (text-only query). In this case, no visual features can be used, thus, a retrieved key-frame cannot be used for initiating a new query.
- Using user-supplied keywords to retrieve shots using the text-ranking algorithm (section 3.4), as well as the high-level face/non-face and motion features. In this case, visual features can be used; thus, despite the absence of sample images for initiating a visual similarity query, a key-frame retrieved using the textual information and the high-level descriptors can be used for submitting a new visual similarity query (section 3.2).
- Using visual examples to start a visual similarity query (section 3.2), and a simple text-processing approach to locate the results of the visual similarity search that are also associated with any user-supplied keywords; this is done so that the latter results receive a higher rank and thus precede any other visual similarity results during presentation to the user. Similarly to the previous query execution option, a key-frame retrieved using all this information can be used for submitting a new visual similarity query.

A Graphical User Interface has been developed to enable the effective combination of all modules in an application that can be globally accessed via the web, allowing the web-based experimentation with the aforementioned query submission schemes.

All the above-described functionalities are illustrated in snapshots of the application, shown in Figure 5

Using this retrieval application and the 24 topics defined for TRECVID 2004, several experiments were conducted. The TRECVID topics are semantics-based topics (e.g. “Find shots of Sam Donaldson’s face - whole or part, from any angle, but including both eyes. No other people visible with him”), that cannot be effectively addressed in a fully automated manner. For these topics, experiments were conducted both using the first option for querying and using any of the last two options, which combine visual similarity search along with text-based search and high-level descriptors. Illustrative results of these experiments, for various topics, are presented in Figure 6. From these it can be seen that, although text is very important for retrieval and can often yield satisfactory results, effectively combining it with visual similarity search and introducing high-level descriptors can significantly improve results. During experimentation with the application, performing visual similarity search using as example images the key-frames retrieved by means of textual search was found to be a particularly effective strategy. The average precision corresponding to each topic is shown in Figure 7.

Given the size of the key-frame collection, an important consideration when building such a retrieval application is the time-efficiency of visual similarity search. Using all the extensions of the MPEG-7 XM described in section 3.2, a visual similarity search in the 33000 key-frames of the TRECVID collection required on the average 2.3 sec. for global image search and 10.3 sec. for region-based search, on a 3GHz P4 PC.

5. CONCLUSIONS

The complete architecture of the developed SCHEMA Reference System was presented in this paper and the various modules integrated with it were outlined.

Using the SCHEMA Reference System, the development of a meaningful application performing indexing and retrieval in the TRECVID 2004 test corpus was reported. In the presented application, the availability of a variety of analysis tools and the combination of various indexing descriptors were employed for alleviating the imperfection of any segmentation algorithm and for countering the variability of image data, respectively, thus greatly improving the usability of the system.

Furthermore, the use of a variety of analysis tools enables their comparative evaluation in terms of their suitability for use in a content-based image/key-frame retrieval system. This, along with the possibility of integrating additional such tools with the SCHEMA reference system, illustrates an additional potential use of it: as a test-bed for evaluating and comparing different algorithms and approaches.

6. ACKNOWLEDGMENT

The authors of this paper, would like to thank the Macedonian Press Agency S.A. for contributing some of the images used in this work for the face/non-face image classifier development.

7. REFERENCES

[1] A. Smeaton, W. Kraaij, and P. Over. The TREC Video Retrieval Evaluation (TRECVID): A Case Study and Status Report. In *Proc. RIAO 2004 - Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, Avignon, France, April 2004.

[2] S.-F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. *IEEE Trans. on Circuits and Systems for Video Technology, special issue on MPEG-7*, 11(6):688–695, June 2001.

[3] MPEG-7 XM software. http://www.lis.ei.tum.de/research/bv/topics/mmdb/e_mpeg7.html.

[4] N. O'Connor, S. Sav, T. Adamek, V. Mezaris, I. Kompatsiaris, T.Y. Lui, E. Izquierdo, C.F. Bennstrom, and J.R. Casas. Region and Object Segmentation Algorithms in the Qimera Segmentation Platform. In *Proc. Third Int. Workshop on Content-Based Multimedia Indexing (CBMI03)*, 2003.

[5] Y. Deng and B.S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(8):800–810, August 2001.

[6] E. Tuncel and L. Onural. Utilization of the recursive shortest spanning tree algorithm for video-object segmentation by 2-D affine motion modeling. *IEEE Trans. on Circuits and Systems for Video Technology*, 10(5):776–781, Aug. 2000.

[7] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Still Image Segmentation Tools for Object-based Multimedia Applications. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(4):701–725, June 2004.

[8] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(8), 2002.

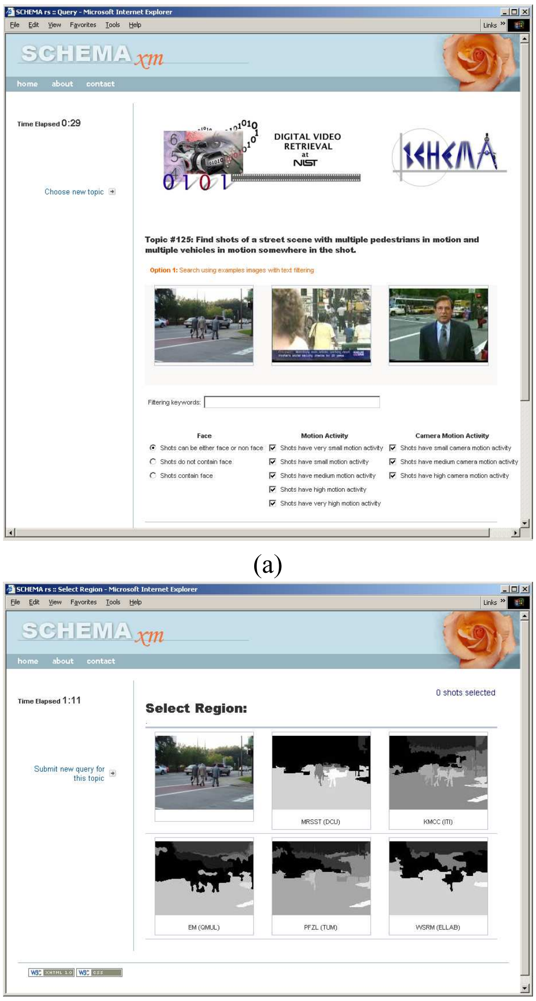
[9] S. Makrogiannis, G. Economou, S. Fotopoulos, and N. G. Bourbakis. Segmentation of Color Images using Multiscale Clustering and Graph Theoretic Region Synthesis. *IEEE Trans. on Systems, Man and Cybernetics: Part A, to appear.*

[10] S. Jeannin and A. Divakaran. MPEG-7 Visual Motion Descriptors. *IEEE Trans. on Circuits and Systems for Video Technology, special issue on MPEG-7*, 11(6):720–724, June 2001.

[11] J.W. Han, L. Guo, and Y.S. Bao. A Novel Image Retrieval Model. In *ICSP Proceedings*, 2002.

[12] J. Luo and A. Savakis. Indoor vs. outdoor classification of consumer photographs using low-level and semantic features. *Proceedings of International Conference on Image Processing*, 2:745–748, 2001.

[13] S.E. Intille and K. Sparck Jones. Simple, proven approaches to text retrieval. *Technical report UCAM-CL-TR-356, ISSN 14762986, University of Cambridge*, 1997.


SCHEMA rs :: Query - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links
SCHEMA xm
home about contact
Time Elapsed 0:29
Choose new topic
DIGITAL VIDEO RETRIEVAL at NIST
SCHEMA

Topic #125: Find shots of a street scene with multiple pedestrians in motion and multiple vehicles in motion somewhere in the shot.

Option 1: Search using examples images with text filtering

Filtering keywords:

Face	Motion Activity	Camera Motion Activity
● Shots can be either face or non face	☑ Shots have very small motion activity	☑ Shots have small camera motion activity
○ Shots do not contain face	☑ Shots have small motion activity	☑ Shots have medium camera motion activity
○ Shots contain face	☑ Shots have medium motion activity	☑ Shots have high camera motion activity
	☑ Shots have high motion activity	
	☑ Shots have very high motion activity	

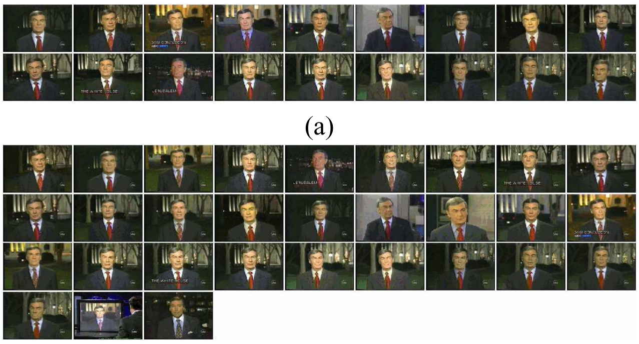
(a)

SCHEMA rs :: Select Region - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links
SCHEMA xm
home about contact
Time Elapsed 1:11
Submit new query for this topic
0 shots selected

Select Region:

(b)

Fig. 5. a) Selecting "Search using example images with text filtering", the user is presented with the topic-specific visual examples supplied by TREC, a field for entering keywords, and several options regarding the high-level features. After entering any keywords and making the high-level descriptor choices using the check boxes, the user can click on any of the example images for specifying an image or region to be used as visual example. b) For supplying visual examples, the user can click on the original image or on any region of any of the employed segmentations.



(a)

(b)

Fig. 6. Illustrative results of the developed application, a) for TRECVID topic 135, "Find shots of Sam Donaldson's face - whole or part, from any angle, but including both eyes. No other people visible with him", using textual information only, and b) using combined textual-visual search.

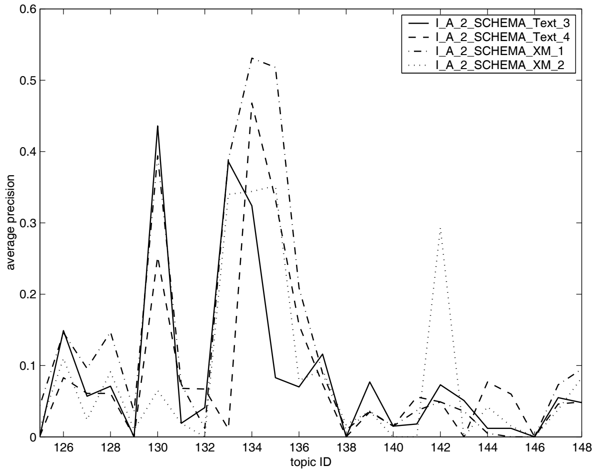


Fig. 7. Average precision for each topic for the four runs submitted to NIST. $I\_A\_2\_SCHEMA\_Text\_3$ and $I\_A\_2\_SCHEMA\_Text\_4$ refer to the partially functional system using text retrieval only, while $I\_A\_2\_SCHEMA\_XM\_1$ and $I\_A\_2\_SCHEMA\_XM\_2$ refer to the fully functional system using both textual and visual similarity. It is clear that the latter is more suitable for video retrieval.