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

Vasileios Mezaris[1,2], *Haralambos Doulaverakis*[2], *Stephan Herrmann*[3], *Bart Lehane*[4], *Noel O'Connor*[4], *Ioannis Kompatsiaris*[2], *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, Greece
[3]Institute for Integrated Systems, Munich University of Technology, Munich D-80290, Germany
[4]Centre for Digital Video Processing, Dublin City University, Ireland

ABSTRACT

In this paper, the most recent version of the system developed by the SCHEMA NoE, termed SCHEMA Reference System, is presented. The Reference System adopts a module-based, expandable architecture, with well defined interfaces between different modules, facilitating the ongoing expansion of the system based on researchers' contributions. The proposed system employs the MPEG-7 XM (MPEG-7 eXperimentation Model) along with extensions developed specifically for the system to improve functionality and efficiency. In addition, the system supports high level descriptors 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

One of the objectives of the SCHEMA Network of Excellence is the design and implementation of a Reference System for content-based information retrieval. The goal of this activity is to develop a software instantiation of a prototypical retrieval system that can be used by other researchers in the field. The benefit of this activity for other researchers in the community is that their own systems could be benchmarked against the Reference System, thereby facilitating more rigorous evaluation of systems. Alternatively, the system could be used as an integration platform for specific component technologies, allowing researchers to evaluate their algorithms in the context of a complete system without having to build such a system themselves. For industry-based researchers the system provides a technology demonstrator of the research currently maturing in research laboratories.

This paper focuses on a description 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 3.3.3). The results obtained using an application 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.

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

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 system combines five different analysis modules developed by different SCHEMA partners and affiliated members. In combination with the low-level descriptors extracted using 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 categorization of images into face/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 proposed system uses the MPEG-7 XM [2] (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. An overview of the proposed architecture is illustrated in Figure 1.

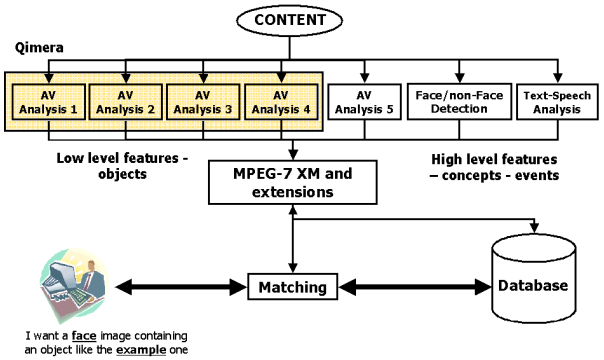

Fig. 1. Overview of the SCHEMA Reference System

3. REFERENCE SYSTEM MODULES

In this section, a description of the different modules comprising the SCHEMA Reference System is presented, starting with the integrated image segmentation algorithms. 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.

3.1. Visual Content Analysis

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 [3], 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 realize the following algorithms:

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

The system can use any module 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 [4]. Illustrative results of the different segmentation modules are presented in Figure 2.

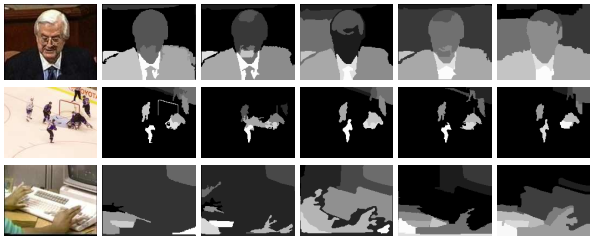

Fig. 2. Sample results of the integrated segmentation modules. The results generated by MRSST, KMCC, EM, PFZL and WSRM are presented in columns 2 to 6, respectively.

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 well-known 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. 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*.

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

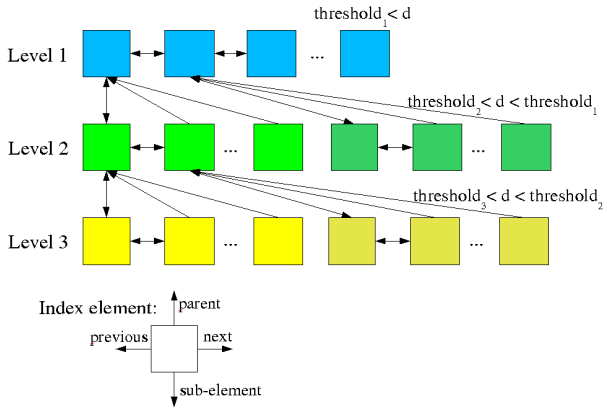

Fig. 3. Index structure with hierarchical element lists.

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. The XM Server uses approaches described in [5] to optimise the search procedure.

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. Considering that the TREC database consists of over 33,000 key-frames, searching through all the key-frames would result in retrieval times not appropriate for interactive video retrieval. 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.

The index set is organized in a multi-level fashion:

- The first level contains the descriptors whose difference exceeds a given threshold, thus each of them is considerably different from each other
- The other levels contain the descriptors which are similar to a descriptor in the first level (their difference is smaller than a threshold). For each depth level of the indexing the similarity threshold value is decreased. In principle, these thresholds influence the number of elements on each level. Thus, they should be adjusted in a way, that a few hundred elements are in each index list.

In this way, a complete index is created, in which the elements in a branch are similar to each other and dissimilar to those of any other branch at the same depth level. As a conclusion the index consists of element lists as shown in Figure 3. For each descriptor there is one index element that has references to the following, previous, parent and sub-element. Thus, it is possible to build double linked lists in horizontal (on one level) and vertical (over hierarchies) directions.

At search time the query descriptor is compared to all elements on the first index level. Then the branch of the best match is followed to refine the search. The process is repeated until the last index level is reached. Finally, the result list is created, thus it can include all matched descriptors. As an extension, the search can follow more than one branch from one index level to the next, to reduce the probability to miss relevant matches.

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).

3.3. High-level and Textual 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" ($\sigma > 32$) [6].

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. 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 [7], [8]. 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 those 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. Experimental results indicate that the proposed region-based classification approach is more accurate that the widely used global image classification.

3.3.3. Textual Information Processing

The text ranking algorithm integrated with the SCHEMA Reference System is the BM25, which incorporates both normalized document length (the associated text for every image/key-frame, in our case) and term frequency [9]. The BM25 algorithm grants higher score to documents, ASR chunks in our case, where query terms/keywords appear more often in them.

4. EXPERIMENTAL RESULTS

The SCHEMA Reference System described in detail in the previous sections was used for building an interactive search and retrieval application, to perform indexing and retrieval in a collection

of over 33,000 key-frames. These correspond to the 64 hours of news video used for the TRECVID 2004 experiments.

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

- Using user-supplied keywords to evaluate shots with the help of the text-ranking algorithm of section 3.3.3 (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, employing the text-ranking algorithm of section 3.3.3 as well as the high-level face/non-face and motion features.
- 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.

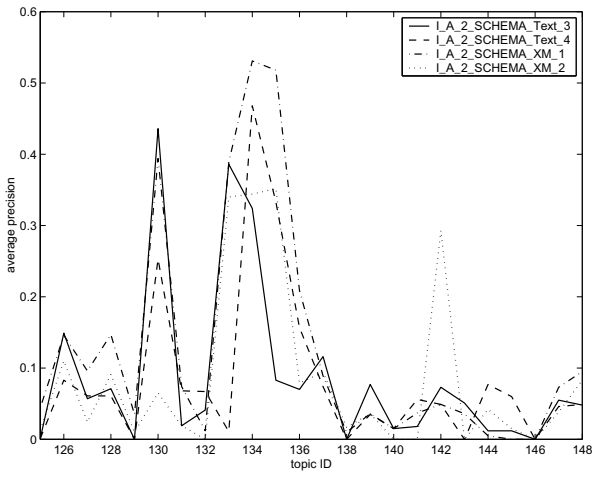

Fig. 4. Average precision for each topic for the four runs submitted to NIST. $I\_A\_2\_SCHEMA\_Text\_3$ and $I\_A\_2\_SCHEMA\_Text\_4$ refer to the partially functional system using text retrieval only, while $I\_A\_2\_SCHEMA\_XM\_1$ and $I\_A\_2\_SCHEMA\_XM\_2$ refer to the fully functional system using both textual and visual similarity. It is clear that the latter is more suitable for video retrieval.

In the latter two cases, visual similarity search can then be performed repeatedly, using one of the query results as example key-frame. Figure 4 shows experimental results using two different variants of the proposed system, one allowing only the first type of queries and one allowing queries of the latter two types. Using these variants and the 24 topics defined for TRECVID 2004, several experiments were conducted. From the results 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 similarity search was found to be a particularly effective strategy.

5. CONCLUSIONS

The complete architecture of the SCHEMA Reference System was presented in this paper. Using the SCHEMA Reference System, the development of a meaningful application performing indexing and retrieval in the TRECVID 2004 test corpus was reported. 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. 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] MPEG-7 XM software, http://www.lis.ei.tum.de/research/bv/topics/mmdb/e_mpeg7.html.

[3] 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.

[4] Y. Deng and B.S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 800–810, August 2001.

[5] V. Mezaris, H. Doulaverakis, R. Medina Beltran de Otalora, S. Herrmann, I. Kompatsiaris, and M. G. Strintzis, "A test-bed for region-based image retrieval using multiple segmentation algorithms and the MPEG-7 eXperimentation Model: The Schema Reference System," in *Proc. 3rd International Conference on Image and Video Retrieval (CIVR), Springer LNCS*, Dublin, Ireland, July 2004, vol. 3115, pp. 592–600.

[6] S. Jeannin and A. Divakaran, "MPEG-7 Visual Motion Descriptors," *IEEE Trans. on Circuits and Systems for Video Technology, special issue on MPEG-7*, vol. 11, no. 6, pp. 720–724, June 2001.

[7] J.W. Han, L. Guo, and Y.S. Bao, "A Novel Image Retrieval Model," in *Proc. 6th International Conference on Signal Processing*, August 2002, vol. 2, pp. 953–956.

[8] 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*, vol. 2, pp. 745–748, 2001.

[9] 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.