# Combining textual and visual information processing for interactive video retrieval: SCHEMA's participation in TRECVID 2004

Vasileios Mezaris, Haralambos Doulaverakis, Stephan Herrmann, Bart Lehane, Noel O'Connor, Ioannis Kompatsiaris, and Michael G. Strintzis.

**Abstract**

In this paper, the two different applications based on the Schema Reference System that were developed by the SCHEMA NoE for participation to the search task of TRECVID 2004 are illustrated. The first application, named "Schema-Text", is an interactive retrieval application that employs only textual information while the second one, named "Schema-XM", is an extension of the former, employing algorithms and methods for combining textual, visual and higher level information. Two runs for each application were submitted, I_A_2_SCHEMA-Text_3, I_A_2_SCHEMA-Text_4 for Schema-Text and I_A_2_SCHEMA-XM_1, I_A_2_SCHEMA-XM_2 for Schema-XM. The comparison of these two applications in terms of retrieval efficiency revealed that the combination of information from different data sources can provide higher efficiency for retrieval systems. Experimental testing additionally revealed that initially performing a text-based query and subsequently proceeding with visual similarity search using one of the returned relevant keyframes as an example image is a good scheme for combining visual and textual information.

**Keywords** : *image segmentation; region-based image and video indexing; MPEG-7 XM; high-level features; multi-modal queries; TRECVID.*

## I. Introduction

The Schema Network of Excellence participated in the Search Task with two interactive retrieval applications based on the Schema Reference System. 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 and reported in detail in [1].

The Schema Reference 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 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. All the aforementioned functionalities have been combined under a common Graphical User Interface, built using web technologies. An overview of the proposed architecture is illustrated in Figure 1.

For participation in the search task of TRECVID 2004 two different applications based on the Schema Reference System were developed, each making use of different modules. The first application, named "Schema-Text", is an interactive retrieval application that employs only textual information while the second one, named "Schema-XM", is an extension of the former, employing algorithms and methods for combining textual, visual and higher level information.

This paper is organized as follows: In section II the different modules integrated with the Schema Reference System are described. In section III the two applications developed for the Schema participation to TRECVID 2004 are presented. Section IV illustrates the experimental procedure which was followed and reports the results that were submitted to NIST for evaluation. Finally, in section V conclusions are drawn.

V. Mezaris and M.G. Strintzis are with the Information Processing Laboratory, Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, and with the Informatics and Telematics Institute (ITI)/Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece. H. Doulaverakis and I. Kompatsiaris are with the Informatics and Telematics Institute (ITI)/Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece. S. Herrmann is with the Institute for Integrated Systems, Munich University of Technology, Munich D-80290, Germany. B. Lehane and N. O'Connor are with the Centre for Digital Video Processing, Dublin City University, Ireland. This work was supported by the EU project SCHEMA "Network of Excellence in Content-Based Semantic Scene Analysis and Information Retrieval" (IST-2001-32795).

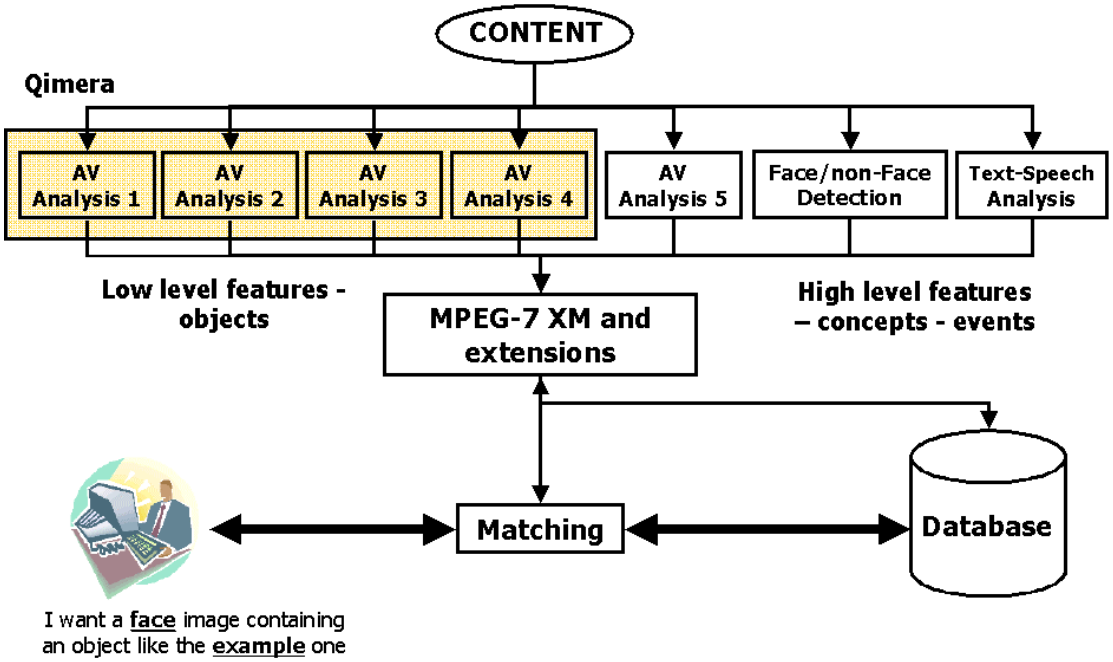



Fig. 1. Overview of the SCHEMA Reference System

## II. Schema Reference System

In this section, a description of the different modules comprising the Schema Reference System is presented. The presentation starts with a description of two (of a total of five) segmentation algorithms that have been integrated into the Reference System. Following that, the indexing and retrieval modules based on the MPEG-7 XM are described and a number of extensions to the XM, developed by SCHEMA partners in order 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.

### *A. Visual Content Analysis*

Two image segmentation modules, out of a total of 5 such algorithms included in the Schema Reference System [2], are briefly presented in this section. 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. These are:

- a Modified Recursive Shortest Spanning Tree algorithm (MRSST);
- a K-Means-with-Connectivity-Constraint algorithm (KMCC);

The segmentations produced by these modules 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].

#### A.1 Modified Recursive Shortest Spanning Tree algorithm

This approach is based on a straightforward extension of the well-known Recursive Shortest Spanning Tree (RSST) algorithm [5]. The original algorithm is modified to avoid merging regions with very different colors. To this end, the RSST is carried out in the HSV color space and is complemented by a second merging stage, in which the creation of large regions is not penalized. This algorithm is explained in detail in [6].

#### A.2 K-Means-with-Connectivity-Constraint algorithm

This segmentation scheme is based on combining a novel segmentation algorithm and a procedure for accelerating its execution. The segmentation algorithm is based on a variant of the K-Means-with-connectivity-constraint algorithm (KMCC) [7], which performs segmentation in a combined color-position-texture feature space so as to produce connected

regions. The acceleration procedure is based on using properly reduced versions of the original images as input and performing partial pixel reclassification after the convergence of the KMCC-based algorithm [7].

### *B. 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*.

The original XM software has certain drawbacks that do not make it appropriate for use in a video retrieval system. Namely these are the time efficiency of the retrieval process, due to binary decoding of the descriptors during each query and the lack of an indexing structure, and the inability to search and combine more than one descriptor at a time during the matching process.

To address these issues several extensions to the original software have been developed and are described in the sequel.

- *MultiImage module extension*: 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.
- *XM Server extension*: 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 software reads in the descriptions from the MPEG-7 bit stream. The query image is then loaded and the query descriptions are extracted. Finally, the query description is compared to all descriptions in the reference database and the most similar images 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/decoded 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 *the XM Server* has been added. XM Server constantly runs as a process in the background (similar to a server). Using this approach the software can accept inputs and perform a search and retrieval process without the need to be executed from the beginning on every query. This greatly reduces search time as the decoding of the binary descriptions are made only once, during the first query. Furthermore if the query image is part of the reference database, the descriptions are are read directly from the database rather than extracted from the image.
- *Indexing module extension*: An indexing structure has been developed that generates clusters of similar images based on the extracted descriptors distance (difference). The result is the grouping of images into groups with high similarity. Therefore when searching is performed only a subset of images are compared thus greatly reducing search time.

### *C. High-level Features*

#### C.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 $\sigma$ 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 [8].

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.

C.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 [9], [10]. 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 classification process is viewed as a typical pattern recognition problem, consisting of training and a testing process.

*D. Textual Information Processing*

The text ranking algorithm integrated with the Schema Reference System is the BM25 algorithm, which incorporates both normalized document length (the associated text for every image/key-frame, in our case) and term frequency [11]. Appropriate values for the parameters used by BM25 have been selected. These values are reported in [11] to produce good results.

## III. Applications developed for TRECVID 2004

Using the methods and modules described above, two applications of the Schema Reference System were developed. In the first application, labelled Schema-Text, the user is only allowed to enter text (keywords) as input. Based on this input the text ranking algorithm of II-D is applied on the ASR text, resulting in a ranked list of hits. The corresponding keyframes, ordered according to rank, are subsequently presented to the user, who is then given the possibility to store the identity of those considered to be relevant results for the given query. This is made possible using a storage structure that mimics the functionality of the shopping cart found in electronic commerce sites. This way, the user can repeat the search using different keywords each time, without loosing relevant shots retrieved during previous runs. No other interaction with the results of the text ranking algorithm is allowed in this application.

In the second application, labelled Schema-XM, a combination of text, visual and high level features is employed for performing retrieval. The user is given two options for initiating a search.

- Under the first option, the user is prompted to select an example image out of those provided by TREC for the given topic, to be used for performing a combined visual and textual similarity search. This is done using as input to the system either the whole image or a region of it, generated using any of the available segmentation algorithms, along with user-supplied keywords. Visual similarity with the supplied example is then evaluated using the MPEG-7 XM, while a simple text-processing procedure is employed to improve the visual similarity ranking of shots that are additionally associated with the query keywords. Results are again presented ordered according to rank and the user is then allowed to store the identity of those considered to be relevant for the given query, as in the first application described above. Further visual similarity searches can subsequently be performed, using as an example image either one of the TREC-supplied ones or an image from the results of a previous search.
- The second option is similar to the first one, with the only difference being the absence of the TREC-supplied example images. Thus, the first query for a given topic can only be initiated using text ranking and the supported high-level features, while further searches can be performed either in a similar fashion or be based on visual similarity, using as an example image one of those retrieved during a previous search.

Under both options, the possibility of fine-tuning the search engine using the high level features is available; for example, a user looking for shots of a particular person could specify that the system retrieves only face shots and/or shots having low motion activity.

As part of the implementation of both applications, a Graphical User Interface was developed to enable their accessibility via the web, allowing different partners of the SCHEMA NoE to experiment with the aforementioned query submission schemes. A timer was also implemented to allow monitoring the time devoted by a user to a given query. All the above-described functionalities are illustrated in snapshots of the application, shown in Figures 2 to 4.

The two applications of the SchemaTREC system, Schema-Text and Schema-XM, where compared against each other in order to assess the impact of enriching text-based retrieval with a combination of visual and high-level features, thus resulting in a retrieval application effectively combining different modalities.

## IV. Experimental Results

In carrying out the interactive search experiments on the TRECVID 2004 topics, a total of 8 users participated, each performing search for 12 search topics: 6 using the Schema-Text application and another 6 using the Schema-XM application. Users were drawn from different research groups around Europe. Based on these experiments, results for 2 runs for each of the two applications were submitted to NIST.

The output of the evaluation conducted at NIST based on the four submitted runs is presented in Figure 5, in terms of average precision for each topic and each run. Average precision foir the purpose of this evaluation is defined as

$$P = \frac{\sum_{i=1}^{N} \frac{i}{Rank_i}}{N} \tag{1}$$

where $N$ is the total number of relevant to the topic shots contained in the test corpus and $Rank_i$ is the rank that the relevant to the topic shot $i$ was awarded. As a result, average precision rewards systems which assign a higher rank to relevant shots, resulting in the latter preceding irrelevant ones when search results are presented to the user. Illustrative results of both developed applications of the Schema Reference System for two of the TRECVID 2004 topics are presented in Figure 6.

Both the evaluation conducted at NIST and the illustrative results of Figure 6 confirm our initial assumptions, namely that the enrichment of text-based querying with visual and high-level features can lead to the retrieval of more relevant shots as compared to text-only searching, thus resulting in improved retrieval performance. Experimental testing additionally revealed that initially performing a text-based query and subsequently proceeding with visual similarity search using one of the returned relevant keyframes as an example image is a good scheme for combining visual and textual information.

## V. Conclusions

Using the SCHEMA Reference System, the development of two applications performing indexing and retrieval in the TRECVID 2004 test corpus was reported. Comparison between the two applications revealed that a multi-modal approach to video retrieval is necessary in order to achieve results better than those of approaches based on a single modality, the latter being text in our experiments. Further analysis of ways for more efficient combination and fusion of results from different analysis tools and sources is required for maximizing the performance of any multi-modal approach.

## References


[1] SCHEMA Public Deliverable 3.1, *First version of the reference system design*, http://www.schema-ist.org/SCHEMA/files/document/30-03-2004/D3.1.pdf.

[2] V. Mezaris, H. Doulaverakis, R. Otalora, S. Herrmann, I. Kompatsiaris, and M. G. Strintzis, "Combining Multiple Segmentation Algorithms and the MPEG-7 eXperimentation Model in the Schema Reference System," in *Proc. IV 2004 - International Conference on Information Visualization*, London, England, July 2004.

[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] 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*, vol. 10, no. 5, pp. 776–781, Aug. 2000.

[6] N. O'Connor, T. Adamek, S. Sav, N. Murphy, and S. Marlow, "Qimera: a software platform for video object segmentation and tracking," in *Proc. Workshop on Image Analysis For Multimedia Interactive Services*, London, UK, Apr. 2003, pp. 204–209.

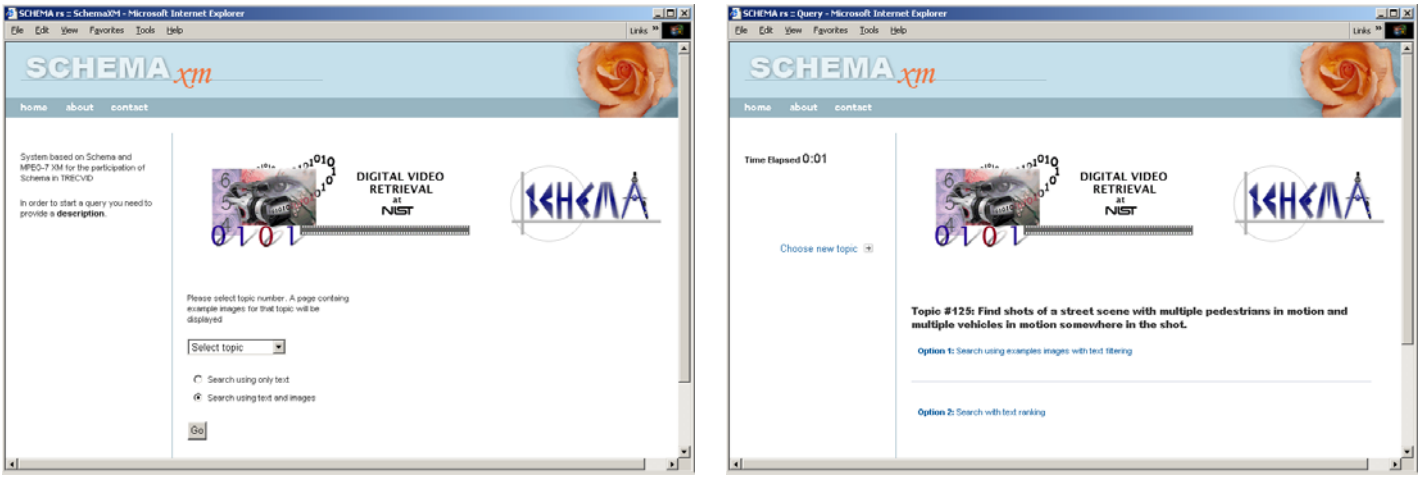
[7] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Still Image Segmentation Tools for Object-based Multimedia Applications," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 4, pp. 701–725, June 2004.

[8] 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.

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

[10] 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.

[11] 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.




(a) (b)

Fig. 2. Query initiation using the developed Graphical User Interface, (a) The user is initially prompted to select a topic and the application to be used for performing the query, namely Schema-Text ("Search using only text") or Schema-XM ("Search using text and images"). In the first case, the complete topic is presented to the user who is then prompted to enter keywords for performing the search, (b) in the case of Schema-XM, its two supported options, "Search using examples images with text filtering" and "Search with text ranking", are presented to the user.


Fig. 3. 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.

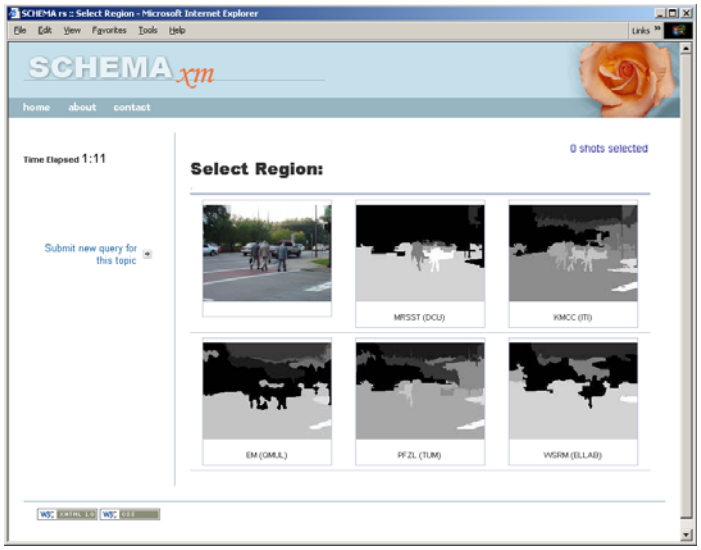



Fig. 4. For supplying visual examples, the user can click on the original image or on any region of any of the employed segmentations. Visual search is then performed using the MPEG-7 XM and its extensions.

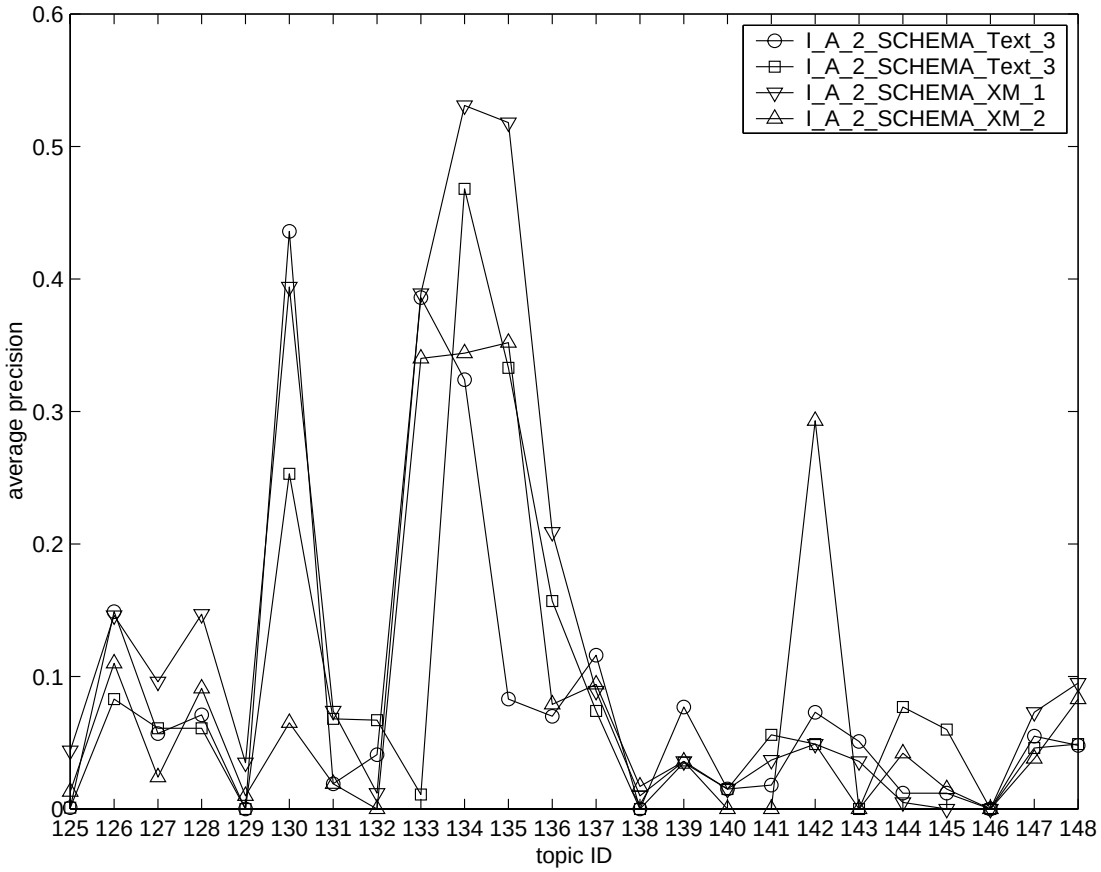



Fig. 5. Average precision for each topic, for the four runs submitted to TREC. The Schema-XM system performs better than the text-only retrieval system in most topics.

(a)


(b)

(c)

(d)

Fig. 6. Illustrative results of the developed application, a) for TRECVID topic 126, "Find shots of one or more buildings with flood waters around it/them", using textual information only, b) using combined textual-visual-high-level search, c) 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 d) using combined textual-visual-high-level search, including the face/non-face classifier.