ADVANCED CONTENT-BASED SEMANTIC SCENE ANALYSIS AND INFORMATION RETRIEVAL: THE SCHEMA PROJECT

E. IZQUIERDO, J.R. CASAS, R. LEONARDI, P. MIGLIORATI, NOEL E. O'CONNOR, I. KOMPATSIARIS AND M. G. STRINTZIS[†]

EC project IST-2000-32795 SCHEMA,
http://www.iti.gr/SCHEMA
E-mail: schema@iti.gr

The aim of the SCHEMA Network of Excellence is to bring together a critical mass of universities, research centers, industrial partners and end users, in order to design a reference system for content-based semantic scene analysis, interpretation and understanding. Relevant research areas include: content-based multimedia analysis and automatic annotation of semantic multimedia content, combined textual and multimedia information retrieval, semantic-web, MPEG-7 and MPEG-21 standards, user interfaces and human factors. In this paper, recent advances in content-based analysis, indexing and retrieval of digital media within the SCHEMA Network are presented. These advances will be integrated in the SCHEMA module-based, expandable reference system.

1. Introduction

The rapid development of innovative tools to create user friendly and effective multimedia libraries, services and environments requires novel concepts to support storage and fast retrieval of huge amounts of digital visual data. Furthermore, the World Wide Web has evolved to a vast distributed system of inter-networked, interactive databases containing text, audio and video stored in a digital form. As a result of almost daily improvements in encoding and transmission schemes, the items of these databases are easily accessible by anyone on the planet. In order to facilitate the rapid retrieval and efficient management of useful information from such multimedia databases, research activities related to content-based multimedia analysis and automatic annotation of semantic multimedia content, combined textual and multimedia information retrieval and semantic-web have evolved along with standards such as MPEG-7 and MPEG-21, user interfaces and human factors.

[†] E. Izquierdo is with the Queen Mary, University of London, J.R. Casas is with the UPC – Technical University of Catalonia, Spain, R. Leonardi and P. Migliorati are with the DEA University of Brescia, Italy, Noel E. O'Connor is with the Dublin City University, Ireland, I. Kompatsiaris and M. G. Strintzis are with the Informatics and Telematics Institute, Greece.

Successful resolution of these matters will allow more efficient and user-friendlier access to all forms of data and will improve data accessibility for all. The diversity and complexity of all these topics requires experts from fields such as signal processing (image-video-audio processing), computer vision, information technology (database design and implementation, multimedia information retrieval), computer networks, human factors engineering and artificial intelligence. Therefore, the aim of the SCHEMA Network of Excellence is to bring together a critical mass of universities, research centres industrial partners and end users, in order to improve the systematic exchange of information and design a reference system for content-based semantic scene analysis, interpretation and understanding.

In this paper, recent advances in content-based analysis, indexing and retrieval of digital media within the SCHEMA network of excellence are presented that are going to be integrated in the project's module-based and expandable reference system. More specifically, after an overview of the current state of the art, two algorithms for knowledge-based inference, one for still images and one for video (soccer game video sequences) are presented. Additionally, two World Wide Web based community-access digital video and image systems are described that could be used as suitable client-server architectures to allow SCHEMA applications to be able to run interactively over the Internet.

2. State of the Art

Many content based indexing and retrieval systems have been developed over the last decade. However, no system or technology has yet become widely pervasive. Most of these systems are currently available for general and domain specific use. These systems fall broadly under four categories: query by content [1, 2], iconic query [3], SQL query [4], and mixed queries [5]. The query by content is based on images, tabular form, similarity retrieval (rough sketches) or by component features (shape, color, texture). The iconic query represents data with 'look alike' icons and specifies a query by the selection of icons. SQL queries are based on keywords, with the keywords being conjoined with the relationship (AND, OR) between them, thus forming compound strings. The mixed queries can be specified by text and as well as icons. All of these systems are based on different indexing structures. In his excellent review of current content based recognition and retrieval systems, Paul Hill [6] describes the most relevant systems in terms of both commercial and academic availability, classified according to the used database population, query techniques and indexing features. Here database population refers to actual process of populating a database. The most popular open available systems are:

QBIC, Photobook, Netra & Netra-V, Virage, Webseek, Islip/Infomedia and ViBE [7].

Cognitively, a predominant feature in video is its higher-level temporal structure. People are unable to perceive millions of individual frames, but they can perceive episodes, scenes, and moving objects. A scene in a video is a sequence of frames that are considered to be semantically consistent. Scene changes therefore demarcate changes in semantic context. Segmenting a video into its constituent scenes permits it to be accessed in terms of meaningful units. A video is physically formed by shots and semantically described by scenes. A shot is a sequence of frames representing continuous action in time and space. A scene is a story unit and consists of a sequence of connected or unconnected shots. Most of the current research is devoted to shot-based video segmentation [8]. Differences between frames can be quantified by pairwise pixel comparisons, or with schemes based on intensity or colour histograms. Motion and dynamic scene analysis can also provide cues for temporal segmentation. A good review of these scene detection schemes is found in [9]. Another approach is proposed by Corridoni and Del Bimbo to detect gradual transitions [10]. They introduce a metric based on the chromatic properties. Ardizzone et al. [11] proposed a neural network approach for scene detection in the video retrieval system JACOB [12]. The approach reported in [13, 14] uses a priori knowledge to identify scenes in a video sequence.

In [15], image annotation or indexing is defined as the process of extracting from the video data the temporal location of a feature and its value. The work by Davis [16] is an example of high level indexing. This approach uses a set of predefined index terms for annotating video. The index terms are organized based on a high level ontological categories like action, time, space, etc. The high level indexing techniques are primarily designed from the perspective of manual indexing or annotation. This approach is suitable for small quantities of new video and for accessing previously annotated databases. To deal with large databases low-level features are needed. Low-level indexing techniques provide access to video based on properties like color, texture etc. One of the pioneering works in this area is by Swanberg et al. [17]. They have presented work on finite state data models for content based parsing and retrieval of news video. Smoliar el al [18] have also proposed a method for parsing news video. Underpinning all indexing techniques in the spatial domain are different processing tasks and methodologies ranging from data base management to low-level image understanding.

3. Knowledge-based inference: from visual features to objects

Segmentation often results insufficient for complex image analysis operations. The paradigm is segmentation applied to object extraction [19]. Objects are

semantic entities by definition, often composed of visually distinguishable parts. This compromises the performance of segmentation algorithms. A number of visual features can be considered to better assess segmentation criteria:

- **Homogeneity**: objects of interest tend to be homogeneous, hopefully with transitions at outer boundaries allowing the definition of contours. Either trivial spatial/temporal or more complex forms of homogeneity such as contour or shape homogeneity can be used as segmentation criteria [20].
- **Compactness (adjacency of parts)**: objects tend to be connected. When objects are composed of different parts, the parts are often spatially linked[c]
- **Regularity (low complexity)**: object shapes and contours usually show some regularity. Most objects of interest present piecewise straight or rounded contour boundaries. Their shape complexity[d] tends to be rather low.
- **Inclusion**: objects may present holes but, most often, smaller parts are included in larger areas. Sometimes inclusion is not complete: larger parts may partially cover smaller parts. In the latter case, smaller parts may still be partially included in the convex hull of larger parts (partial inclusion).
- **Symmetry**: Symmetry abounds in most natural and artificial objects [21, 22]. Segmentation algorithms tend to disregard symmetry for the complexity of analysis. We put forward symmetry as a strong feature for object extraction.

The features listed above define the visual structure of objects and, as such, are related to their physical structure, showing up directly in the visual signal. We call these features "syntactic" features, as opposed to semantic. Like spatial homogeneity, syntactic features are not specific to a particular kind of targeted or modeled objects and, thus, their use for object extraction remains generic, not narrowing the scope of the application domain for segmentation algorithms.

Syntactic features can be found by structure (or syntax) analysis. Structure analysis is based on shapes and spatial configuration of spatially homogeneous regions in the image. Shape and structure are difficult to assess for segmentation criteria at the level of the pixels, before any segmentation has been carried out. Therefore, structure analysis is more conveniently carried out starting from a simple initial segmentation (possibly an over-segmented image). This allows assessing shape and position criteria for the initial regions in order to detect structures formed by sets of simple regions that might be proposed as object candidates.

---

[c] Object compactness is one reason why segmentation is applied to object extraction. Segmentation algorithms produce partitions, i.e. sets of connected (compact) regions.

[d] Shape complexity can be defined as the squared contour length divided by the area.

An example is shown in Figure 1. The proposed features have been analyzed over the initial partition. A merging algorithm progressively structures the regions into sets of quasi-symmetric or partially included regions until the final partition. Observe, how inclusion allows the merging of regions 1 and 2 into the background in Figure 1.d). Partial symmetry with respect to a central element allows the merging of regions 4 and 5 with 3 (the central element). Texture information and transition strength between regions is also considered at each merging. This explains why the last two lower regions are merged within the background, and not with the central element (former regions 3+4+5) what might happen if only symmetry was considered.

Figure 1. Structure analysis using symmetry, compactness, regularity and inclusion: a) initial partition (57 regions), b) 24 regions, c) 8 regions, d) 3 regions e) 2 regions.

4. Sport Content Characterization by Controlled Markov Chains

In this section, a semantic indexing algorithm based on the controlled Markov chain-modelling framework is proposed. The proposed algorithm has been conceived for soccer game video sequences. The problem of automatic detection of semantic events in sport games has been studied by many researchers. In general the objective is to identify certain spatio-temporal segments that correspond to semantically significant events.

In [23], for example, presents a method that tries to detect the complete set of semantic events, which may happen in a soccer game. This method uses the position information of the player and of the ball during the game as input, and therefore needs a quite complex and accurate tracking system to obtain this information. In [24] and [25] the correlation between low-level descriptors and the semantic events in a soccer game have been studied. In particular, in [24], it is shown that the low-level descriptors are not sufficient, individually, to obtain satisfactory results. In [26] the temporal evolution of the low-level descriptors in correspondence with semantic events has been exploited, by proposing an algorithm based on a finite-state machine. This algorithm gives good results in terms of accuracy in the detection of the relevant events, whereas the number of false detections remains still quite large.

In this work, a semantic video indexing algorithm using controlled Markov chains to model the temporal evolution of low-level descriptors was studied. Certain low-level descriptors were chosen, which represent the following

characteristics: (i) lack of motion, (ii) camera operations (pan and zoom parameters), and (iii) presence of shot-cuts. It is supposed that each semantic event takes place over a two-shot block and that it can be modelled by a controlled Markov chain.

Specifically, 6 models were considered denoted by A, B, C, D, E, and F, where model A is associated to goals, model B to corner kicks, and models C, D, E, F describe other situations of interest that occur in soccer games, such as free kicks, plain actions, and so on. On the basis of the derived six Markov models, one can classify each pair of shots in a soccer game video sequence by using the maximum likelihood criterion.

The performance of the proposed algorithm have been tested considering about 2 hours of MPEG2 sequences containing more than 800 shot-cuts, determined using the algorithm described in Section 2. The sequences contain 9 goals and 16 corner kicks. The obtained results are the following: 8 goals out of 9, and 10 corner kicks out of 16 are detected. The number of false detections could seem quite relevant. However, these results are obtained using motion information only, so using other type of media information could probably reduce these false detections. Therefore, further work needs to be done in order to improve its performance by considering other descriptors, related for example to audio information.

5. Físchlár: a suite of online multi-media information retrieval systems

The Centre for Digital Video Processing (CDVP) in Dublin City University has developed a Web-based community-access digital video system called Físchlár [26]. The system allows recording, analysis, browsing, searching and playback of content from 8 terrestrial television channels. It stores over 300 hours of MPEG-1 encoded content at any time, supports over 200 simultaneous video streams and has over 1400 users on campus. The system is used as an experimentation platform for testing and evaluating automatic audio-visual indexing tools and structured design methods for user interfaces. A version of the system designed specifically for broadcast news has been deployed for educational use by journalism students, whereas a version providing medical content is used by nursing students.

The most recent version of the system was developed as part of the CDVP's work in the video track of the Text Retrieval Conference (TREC) – a U.S. initiative to benchmark IR [27]. TREC provided a test corpus, selected a set of audio-visual features to be extracted and specified a set of queries. Features included indoor scenes, outdoor scenes, presence of a face, presence of a group of people, cityscape, landscape, text overlay, speech, instrumental sound, monologue. Targeting three of these, the CDVP developed automatic

approaches for speech/music discrimination and rhythm detection [28] and face detection. The results of participants' different feature extraction tools were made available to all participants in MPEG-7 format. A Físchlár interface for searching the test corpus using all available TREC features was developed and is illustrated in Figure 2. It allows formulation of a query based on selecting features from the available set and browsing and ranking of results based on different combinations of features.

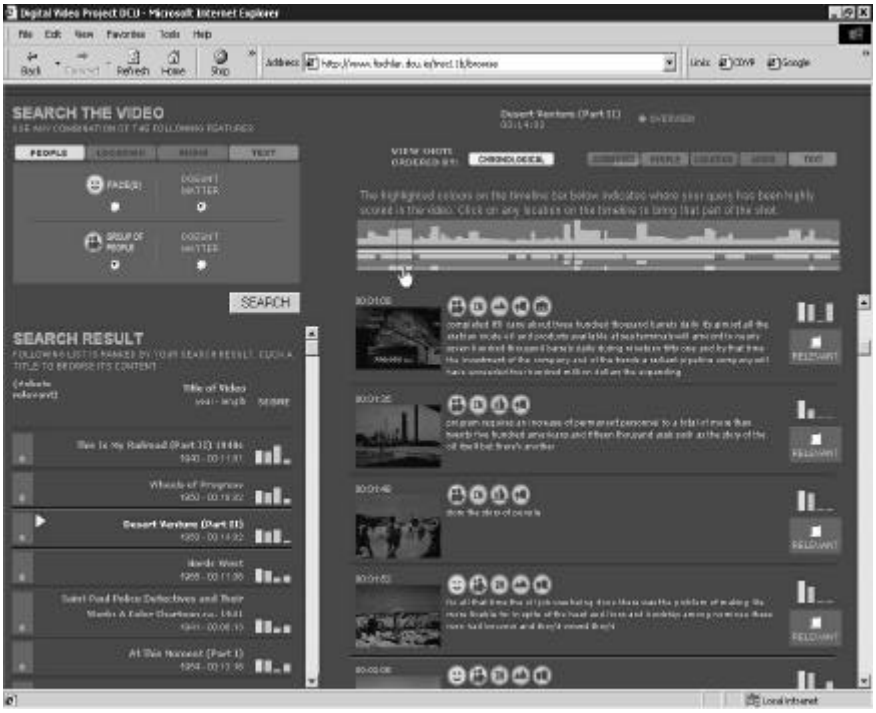

Figure 2 - The Fischlar-TREC search and retrieval interface

6. A World Wide Web Region-Based Search Engine

ISTORAMA[e] is a region-based color indexing and retrieval system for the Internet. As a basis for the indexing, a novel K-Means segmentation algorithm is used, modified so as to take into account the assumed region coherence [29], [30]. Based on the extracted regions, characteristic features are estimated using color, texture and shape information. This algorithm is integrated into a

[e] http://uranus.ee.auth.gr/Istorama

complete content-based search engine system using web and database technologies [31]. An important and unique aspect of the system is that, in the context of similarity-based querying, the user is allowed to view the internal representation of the submitted image and the query results. More specifically, in a querying task, the user can access the regions directly in order to see the segmentation of the query image and specify which aspects of the image are central to the query.

The overall system is split into two parts: (i) the on-line part and (ii) the on-line or user part. In the on-line part, Information Crawlers, implemented entirely in Java, continuously traverse the WWW, collect images and transfer them to the central Server for further processing. Then the image indexing algorithms process the image in order to extract descriptive features. Based on the extracted by the modified K-means algorithm [29], [30] regions, characteristic features are estimated using color, texture and shape/region boundary information. The characteristic features along with information regarding the images such as the URL, date of transaction, size and a thumbnail are then stored in the database. For the database access and management, the MySQL database management system was used.

In the on-line part, a user connects to the system through a common Web Browser using the HTTP protocol. The user can then submit queries either by example images or by simple image information (size, date, initial location, etc). The query is processed by the server and the retrieval phase begins; the indexing procedure is repeated again for the submitted image and then the extracted features are matched against those stored in the database using an SQL query. The results containing the URL as well as the thumbnail of the similar images are transmitted to the user by creating an HTML page with use of PHP (recursive acronym for PHP: “Hypertext Preprocessor"). The results are ranked according to their similarity to the submitted image.

In order to evaluate the performance of the system a variety of queries using a set of 3,500 images were performed. In most cases, the retrieved images match the selection criteria of the user. A comparison to global histogram was made and it was seen that the global histogram matching performs much worse than that of the described system, with average lower precisions in all image categories.

References

1. S.K. Chang and T. Kunii, “Pictorial Database Systems,” IEEE Computer, Ed. S.K. Chang, November 1981, pp. 13-21.
2. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkhani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, ``Query

by Image and Video Content: The QBIC System," IEEE Computer, Vol. 28, No. 9, September 1995, pp. 23-32.

3. A.D. Bimbo, M. Campanai, and P. Nesi, ``A Three-Dimensional Iconic Environment for Image Database Querying," IEEE Trans. on Software Engineering, Vol. 19, No. 10, October 1993, pp. 997-1011.
4. J.A. Orenstein, and F.A. Manola, ``PROBE Spatial Data Modeling and Query Processing in an Image Database Application," IEEE Trans. on Software Engineering, Vol. 14, No. 5, pp. 661-629, May 1988.
5. G. Ahanger, D. Benson, and T.D.C. Little, ``Video Query Formulation," Proc. IS&T/SPIE, Conference on Storage and Retrieval for Image and Video Databases, Vol. 2420, February 1995, pp. 280-291.
6. P. Hill, 'Review of current content based recognition and retrieval systems', Technical report 05/1, Virtual DCE.
7. J-Y Chen, C.A. Bouman, and John Dalton. "Similarity pyramids for browsing and organization of large image databases", Proc. SPIE, vol. 3656, pp 144-154, 1999.
8. F. Beaver. Dictionary of Films Terms. Twayne Publishing, New-York, 1994.
9. G. Ahanger and T.D.C. Little. A survey of technologies for parsing and indexing digital videos. Journal of Visual Communication and Image Representation, 7:28—43, March 1996.
10. J.M. Corridoni and A. Del Bimbo. Structured digital video indexing. In ICPR '96, pages 125—129, 1996.
11. E. Ardizzone, G.A.M. Gioiello, M. La Cascia, and D. Molinelli. A real-time neural approach to scene cut detection. In SPIE Storage and Retrieval for Image and Video Databases IV, 1996.
12. M. La Cascia and E. Ardizzone. Jacob: Just a content-based query system for video databases. In ICASSP'96, 1996.
13. D. Swanberg, C.-F. Shu, and R. Jam. Knowledge guided parsing in video databases. In SPIE vol.1908, pages 13—21, 1993.
14. H. Zhang, Y. Gong, S.W. Smoliar, and S.Y. Tan. Automatic parsing of news video. In International Conference on Multimedia Computing and Systems, pages 45—54, 1994.
15. A Survey of Technologies for Parsing and Indexing Digital Video, Boston University, http://hulk.bu.edu/pubs/papers/1995/ahanger-jvcir95/TR-11-01-95.html
16. Deborah Swanberg, Chiao-Fe Shu, and Remesh Jain. "Knowledge guided parsing in v ideo databases." Electronic Imaging: Science and Technology, San J ose, California, February 1993. IST/SPIE.
17. Stephen W Smoliar, HongJiang Zhang, and Jian Hua Wu. "Using frame technology to manage video." In Proc. of the Workshop on Indexing and

Reuse in Multimed ia Systems. American Association of Artificial Intelligence, August 1994
18. Arun Hampapur, Ramesh Jain and Terry E Weymouth, "Feature Based Digital Video Indexing"
19. S-F Chang, et al. Semantic Visual Templates: Linking Visual Features to Semantics. International Conference on Image Processing (1998)
20. B. Johanson. Multiscale Curvature Detection in Computer Vision. Thesis No. 877, Department Electrical Eng., Linköping University, Sweden (2001)
21. P. J. van Otterloo, A Contour-Oriented Approach to Shape Analysis. Prentice Hall, London (1991).
22. C. Sun and D. Si, "Fast Reflectional Symmetry Detection Using Fourier Transform", Journal of Real-Time Imaging, vol.5, no.1, p.63-74 (Feb 1999).
23. V. Tovinkere, R. J. Qian, “Detecting Semantic Events in Soccer Games: Toward a Complete Solution”, Proc. ICME'2001, pp. 1040-1043, August 2001, Tokyo, Japan.
24. A. Bonzanini, R. Leonardi, P. Migliorati, “Semantic Video Indexing Using MPEG Motion Vectors", Proc. EUSIPCO'2000, pp. 147-150, 4-8 Sept. 2000, Tampere, Finland
25. A. Bonzanini, R. Leonardi, P. Migliorati, “Event Recognition in Sport Programs Using Low-Level Motion Indices”, Proc. ICME'2001, pp. 920-923, August 2001, Tokyo, Japan.
26. N. O'Connor, et al, “Físchlár: An On-line System for Indexing and Browsing of Broadcast Television Content”, Proc ICASSP 2001, Salt Lake City, UT, 7-11 May 2001.
27. A. Smeaton et al, “The TREC2001 Video Track: Information Retrieval on Digital Video Information”, ECDL 2002 - European Conference on Research and Advanced Technology for Digital Libraries. Rome, Italy, 16-18 September 2002.
28. R. Jarina et al, “Rhythm Detection for Speech-Music Discrimination in MPEG Compressed Domain”, Proc. DSP 2002 - 14th International Conference on Digital Signal Processing, Santorini, Greece, 1-3 July 2002.
29. I. Kompatsiaris and M.G. Strintzis, "Spatiotemporal Segmentation and Tracking of Objects for Visualization of Videoconference Image Sequences", IEEE Trans. on Circuits and System for Video Technology, vol. 10, no. 8, December 2000.
30. N. V. Boulgouris, I. Kompatsiaris, V. Mezaris, D. Simitopoulos, M. G. Strintzis, "Segmentation and Content-based Watermarking for Color Image and Image Region Indexing and Retrieval", EURASIP Journal on Applied Signal Processing, April 2002.

31. I. Kompatsiaris, E. Triantafillou and M. G. Strintzis, "Region-Based Color Image Indexing and Retrieval", 2001 International Conference on Image Processing (ICIP2001), Thessaloniki, Greece, October 7-10, 2001.