A MPEG-7 compliant Video Management System: BilVMS

Ersin Esen[1,2], Özgür Önür[1,2], Medeni Soysal[1,2], Yagiz Yasaroglu[1,2], Serhat Tekinalp[1] and A. Aydin Alatan[1,2]

[1]Department of Electrical-Electronics Engineering, M.E.T.U.,
[2]TÜBITAK BILTEN,
Ankara, 06531 TURKEY

In the recent years, there has been a growing interest towards the management of multimedia data. Recently, ISO MPEG organization has finished establishing a new standard, MPEG-7, for describing audio-visual data for interoperable indexing, searching and browsing purposes. Following this standard, a state-of-the-art video management system has been designed and implemented. The system is capable of temporally segmenting video into shots, as well as obtaining a semantically meaningful group of shots, i.e. scenes. The scene decomposition is achieved using a HMM-based formulation by multimodal features. Keyframes are used as shot representatives and their visual descriptions are utilized to make similarity queries. Moreover, these low-level descriptors are also used to reach a number of semantic visual classes using support vector machines. Finally, automatic detection of human faces via skin color filtering and videotext recognition increase the indexing capabilities of the BilVMS system.

1. Introduction

In the recent years, there has been a growing interest towards the management of multimedia data. Recently, ISO MPEG organization has finished establishing a new standard, MPEG-7 [1], for describing audio-visual data for interoperable indexing, searching and browsing purposes. MPEG-7 standardizes how to describe any audio-visual data, as well as some higher level relations, between these descriptions.

2. System Overview

The block diagram of the system is shown in Figure 1. There are five main components: Input module, Video analyzer, Image analyzer, Streaming and database servers, and Client. The input module is capable of capturing data from a video card and storing in various formats and bit-rates. It is also possible to store the video in one of the streaming formats for a lower bit-rate and annotating the video based on MPEG-7 MDS[1]. Video analyzer and Image analyzer are described in Sections 3 and 4, respectively. The Client can be reached at http://vms.bilten.metu.edu.tr

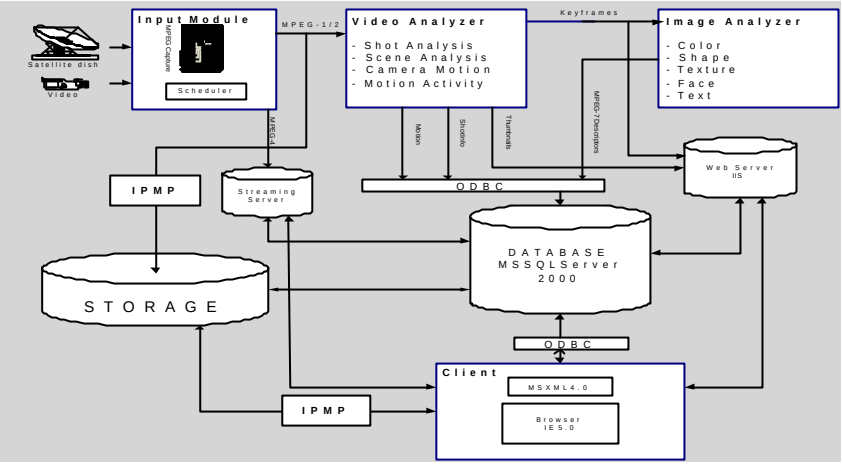

Figure 1. System block diagram for BilVMS

3. Video Analysis

3.1. *Shot Boundary Detection, Keyframe Selection and Scene Analysis*

Shot boundary detection is achieved by combining the histogram and intensity differences between consecutive frames using 2-means clustering. This off-line algorithm [5] merges points, which result from consecutive frame pairs, around two separate clusters, as shot and non-shot boundaries. The simulation results on MPEG-7 test set give a precision and recall rate, as %94 and %92, respectively.

For better summarization, the shots should be clustered together according to their scene content, which is called as scene analysis [4]. As a first step, the compressed input bit-stream is demultiplexed into its audio and video components. The shot information is used in audio analysis to determine whether the audio track of that shot contains speech (T), silence (S) or music (M) content [4]. Shot segmented video is further examined to detect faces within frames for each shot (Section 4.3). If there exists a face in a shot, it is labeled as (F), if not, no-face (N). Similarly, a location analysis can be achieved to check whether the observed scene (location) is changed (C) or remained unchanged (U) between consecutive shots [4]. The Token Generator simply combines the output of three analysis kernels to obtain one token for each shot. Finally, the model parameters (state change and state output probabilities) are used to obtain the best state sequence corresponding to input sequence by dynamic programming (i.e. Viterbi algorithm). For the proposed scene analysis approach, dialogue scenes are segmented using a circular 2-state HMM with dialogue and non-dialogue states. Using this model, it is possible to parse the consecutive input tokens (i.e. shots) between these two states (scenes) for a better

summarization is possible. The simulation results on MPEG-7 test set indicate a performance around %85 for correct scene classification.

3.2. *Motion Descriptors*

The motion descriptors of the presented system are determined by MPEG-7 visual descriptors [1]. While the motion activity descriptor measures the overall motion content in a shot, the camera motion descriptor enables indexing any camera activity. These two descriptors are capable of explaining any temporal action in the scene.

4. Image Analysis

4.1. *Color and Texture Descriptors*

The color and texture descriptors of the presented system are determined by MPEG-7 visual descriptors [1]. These descriptors are dominant color, scalable color, color layout, color structure, homogenous texture and edge histogram. These descriptors have different properties and can be useful for different applications.

4.2. *Semantic Classifications using Support Vector Machines*

The ultimate goal for visual indexing is to reach semantic information from low-level video descriptors. Although, this goal is quite ambitious for general scenes, for a specific visual class, this can be achieved via supervised classification.

Images are assigned to a number of semantic classes according to their MPEG-7 color and texture features. Currently, the selected semantic classes are *Indoor-Outdoor*, *Crowd* and *Soccer*. In order to classify these classes, the utilized MPEG-7 descriptors are edge histogram (vertical components), homogeneous texture and color layout, respectively [1]. The classifier is a support vector machine (SVM) [6] which is trained by supervised data before the test stage. After classification using SVM, the resulting classes can be retrived using the “thin client” of BilVMS. An average %82 detection performance is achieved for different sets of data for *Crowd* and *Indoor-Outdoor* tests.

4.3. *Videotext Detection & Recognition and Face Detection*

The detection of videotext regions using texture analysis is considered as a classical supervised pattern recognition problem. The classifier is selected as a 3-layer single output feed-forward neural network, which has the well-known capability of discriminating linearly inseparable classes. The feature vector to be classified is extracted by using wavelet transform. This vector is fed to the network and the network responds to this input by giving an output indicating the type of the input. After exploiting the textural discriminatory, the image is further analyzed to remove large and very small constant intensity regions. The resulting image is locally thresholded using iterative thresholding method [2].

After a final heuristic to check the spatial distribution of the regions, the resulting image is forwarded to optical character recognition stage (OCR).

The system is capable of character recognition rate up to 59%, which is quite reasonable for most purposes. It should be emphasized that text box detection rate is almost perfect for the proposed system and the character recognition rate, which depends on the OCR performance, for the manually segmented text boxes is found as 65% for the utilized OCR and image resolution (352x288), which should be an upper bound for the proposed system.

Face detection is a quite mature topic with diverse solutions [3]. Although, there are also sophisticated algorithms to validate the existence of a face for a skin-colored region, some simple heuristics still help to detect faces. Using heuristic rules such as aspect ratio, location and area of the skin-colored regions, a shot is labeled as (F) or (N). Note that these heuristics can be increased (e.g. circularity of the region, detection of eyes/mouth, etc.) while compromising from complexity.

More sophisticated methods based on the appearance of the (gray level) face regions is also tested. Although, there is some computational burden, the classification between face and no-face multiple classes using Fisher Linear Discriminant is found out to be quite useful [3].

5. Conclusions

The presented system is capable of indexing visual data in terms of MPEG-7 descriptors. In this way, automatic retrieval of visual similarities and classification of some semantic concepts become possible. While, videotext recognition capability adds important information about the content to the database, the scene-based summarization is a versatile tool for better comprehensiveness of the summary.

References

1. ISO/IEC JTC1/SC29/WG11/W3703 MPEG-7 Multimedia Content Description Interface – Part 3 Visual, October 2000.
2. S. Tekinalp, “Detecting and Recognizing Text from Video Frames”, M.S. Thesis, Middle East Technical University, September 2002.
3. E. Taslidere, “Face Detection using a Mixture of Subspaces”, M.S. Thesis, Middle East Technical University, September 2002.
4. A. A. Alatan, A. N. Akansu and W. Wolf, “Multi-modal Dialogue Scene Detection Using Hidden Markov Models for Content-based Multimedia Indexing,” *Kluwer Academics, International Journal on Multimedia Tools and Applications*, 14, 2001, pp137-151.
5. M. Naphade, R. Mehrotra A. Mufit Ferman, J. Warnick, T. S. Huang, and A.M. Tekalp, "A High-Performance Shot Boundary Detection Algorithm Using Multiple Cues,” in *Proc. IEEE ICIP'98*, Chicago, Ill., Oct. 1998.
6. S. Tong and E. Chang, “Support Vector Machine Active Learning for Image Retrieval,” Proceedings of the 9th ACM international conference on Multimedia, Ottawa, Canada, 2001, pp 107-118.