Automatic Video Management System Using Face Recognition and MPEG-7 Visual Descriptors

Jae-Ho Lee

ABSTRACT—The main goal of this research is automatic video analysis using a face recognition technique. In this paper, an automatic video management system is introduced with a variety of functions enabled, such as index, edit, summarize, and retrieve multimedia data. The automatic management tool utilizes MPEG-7 visual descriptors to generate a video index for creating a summary. The resulting index generates a preview of a movie, and allows non-linear access with thumbnails. In addition, the index supports the searching of shots similar to a desired one within saved video sequences. Moreover, a face recognition technique is utilized to personal-based video summarization and indexing in stored video data.

Keywords—Face recognition, multimedia document authoring, MPEG-7.

I. Introduction

The popularity of digital equipment and the Internet allows us to come into contact with a large amount of video data. The sheer amount of data is becoming increasingly difficult to handle on conventional systems. Utilization of MPEG-7 is a reasonable approach to describe and manage multimedia data [1]. To this end, there has been some research done on the use of MPEG-7 in broadcasting content applications. A. Yamada and others have built a visual program navigation system that uses an MPEG-7 color layout descriptor [2]. K.O. Kang and others designed a metadata broadcasting system [3], T. Walker proposed a system for content-based navigation of television programs based on MPEG-7 [4], and N. Dimitrova and others applied their own metadata [5]. The system that Walker presented uses standard MPEG-7 description schemes (DS) to describe television programs. T. Sikora applied the MPEG-7 descriptor for the management of multimedia databases [6]. Also, A. Divakaran and others presented a video summarization technique using cumulative motion activity based on compressed domain features extracted from motion vectors [7]. The above systems have shown reasonable results in specific genres such as news or sport games.

In this paper, a video management system that generates summarized video efficiently is introduced. The summary information generated by the presented tool provides users with an overview of video content and guides them visually to move to a desired position quickly within a video. Also, the tool makes it possible to find shots similar to a queried one or a particular face, which is useful for editing different video streams into a new one. In this work, MPEG-7 visual descriptors are only used to segment a video into shots, to summarize a video, and to retrieve a scene of interest.

II. System Overview

The functions of the developed system consist of mainly four parts:

a. Generation of an overview of video contents: to find a desired position in video data.
b. Query by example or face: to find similar shots to a queried one in a large amount of video data.
c. Nonlinear editing: to support simple editing based on the summarized video.
d. Actor-based indexing: to find scenes which contain a queried person in a video file.

Figure 1 shows an overview of the implemented automatic video management system. In this system, face information works for actor-based video indexing. Actually, most

Manuscript received July 06, 2005; revised Sept. 30, 2005.
Jae-Ho Lee (phone: +82 42 860 5428, email: jhlee3@etri.re.kr) is with Digital Content Research Division, ETRI, Daejeon, Korea.

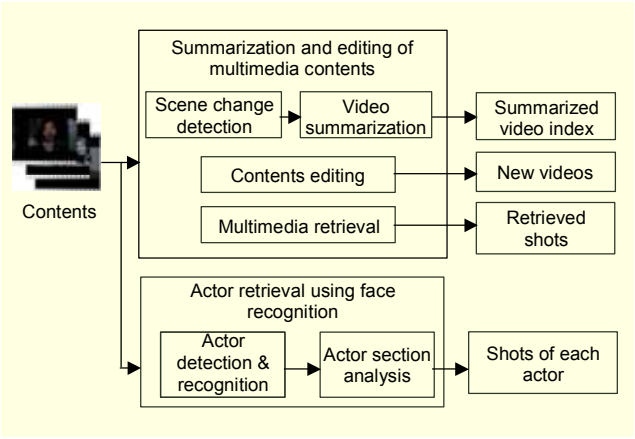

Fig. 1. Overview of the video management system.

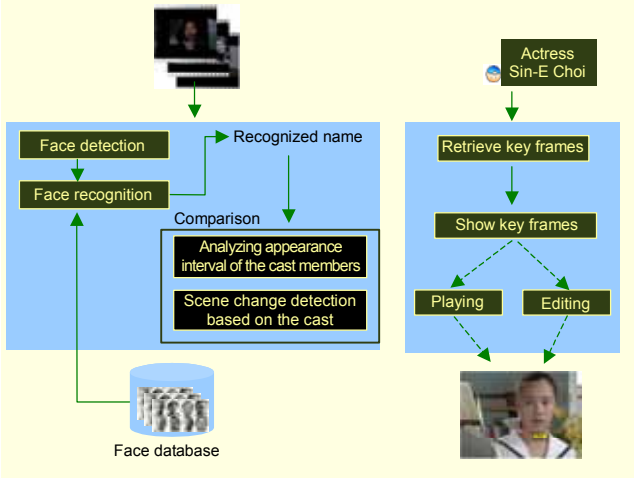

Fig. 2. Employment diagram of face information in the system.

broadcasting dramas or films are focused on the shot of an actor. If face information is adopted properly, the user can access or retrieve his or her favorite scene. Actor-based indexing makes this approach possible. Figure 2 illustrates how face information is utilized in the implemented video management system. In this system, face detection is processed every five frames to realize a real-time system. If a face is detected, the detection procedure is halted and the detected face is reiterated as the representative face of that shot.

When a face is detected, it is recognized using a stored face database. If a face is not classified, it can be registrated as an "unknown" in the face database. This information can be edited by a supervisor later. With the recognized face, a user can access, review, and edit the video data easily. The face information can also be used in scene change detection if the faces are recognized well. Figure 3 shows the graphic user interface of the developed system.

In Fig. 3, the left interface is for video display and setting the processing options. The right side displays the content information by each shot or scene. The user can access his desired shots with the displayed information, and retrieve or edit the videos with this information. Basically, the face information can be useful for a user to access the desired position in video data such as in a drama or movie because the actor is the main component of these genres.

Figure 4 illustrates the time line of each actor when the actor tab shown in Fig. 3 is selected. The upper-right region is the list of actors in the experimental video data. This actor information, which is automatically generated, helps the user to find his or her desired shots and edit with these scenes.

The system generates actor-based indexing in real-time. In a face dataset, all face images are normalized manually, and this generates higher recognition results than real video data. However, the faces were normalized with the proposed Active Shape Model (ASM) in the eye region in our experimental results. Several TV drama clips were selected to estimate the performance of the proposed method by Korea Broadcasting System, which is trying to develop an automatic broadcasting management system [8]. The recognition performance has a

Fig. 3. GUI of overall system.

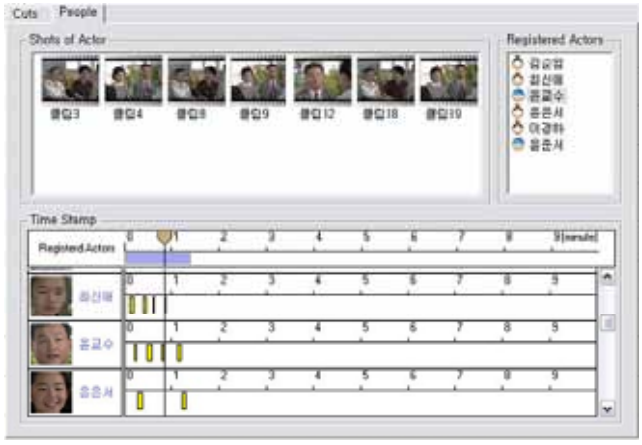

Fig. 4. Actor-based video indexing.

success rate of 93.45%. In the face detection stage, the Haar-like/AdaBoost method is used to make a real-time system. The proposed Haar-like/Linear Discriminant Analysys (LDA) face recognizer is adapted in this system.

The recognition rate of this experiment is relatively high. However, one false recognition can lead to a severe affection in the indexing system. To avoid this problem, a re-adjustment of the index must be considered. This is an approach that uses the similarities of detected scenes. In this system, the visual features of all key frames are saved for retrieval and indexing. If the visual features of detected scenes are similar and face recognition results are different in neighbored scenes, the most frequent face information will be used to correct the information for the face in question.

III. Overview of the Techniques

If a video is inserted into the system, the scene change detection process is activated using MPEG-7 visual descriptors [9]. After the scene change detection step, a video summarization process follows. The algorithm proposed by Yeung is used in this step [10]. In our experiments on video data, the face is localized using the boosted cascaded classifier method of Viola and Jones [11]. The face location is predicted approximately using a suitable local detector, which can be seen in Fig. 3.

After localization of a face, a registration process for face recognition is required to achieve a higher performance. This process is usually called face normalization or face registration. In appearance-based face recognition, which is a traditional face recognition approach, this normalization process is one of the most influential problems. An ASM and active appearance model (AAM) are normally used to solve these registration problems in the latest researches [12]. To reduce the processing time, the ASM was adopted on the eye region, not the whole face region. This solution supports quick and simple processing for face normalization, and this short processing is useful for a huge amount of video data processing.

In our experiments, only sixteen landmarks are utilized to fit the model to the new eye image. When a face is localized, the eye region can be estimated suitably with the structure information of the face. Then, the initial point of an eye ASM is located on that position, and the new eye shape is estimated after several iterations. Figure 5 illustrates the results of the proposed method to normalize a face for the further face recognition process.

Fig. 5. Result of normalization using proposed eye ASM.

As shown in Fig. 5, the degrees of rotation and scale can be estimated in the new image when the eye model is fitted. Therefore, the face can be rotated and scaled to the previously defined position that is proper to face recognition. In the experimental results, the face recognition ratio with this proposed normalization is addressed to prove that the eye ASM is a suitable methodology in the aspects of time complexity and performance of normalization.

In this experiment, the MPEG-7 face data set has been utilized. The MPEG-7 face data set contains various commonly used face data sets, which have been constructed from some universities or institutes, to produce an independent face recognition descriptor on the data sets. Among the MPEG-7 face data, 2000 face images were selected randomly. The results of the conventional approaches were verified in MPEG-

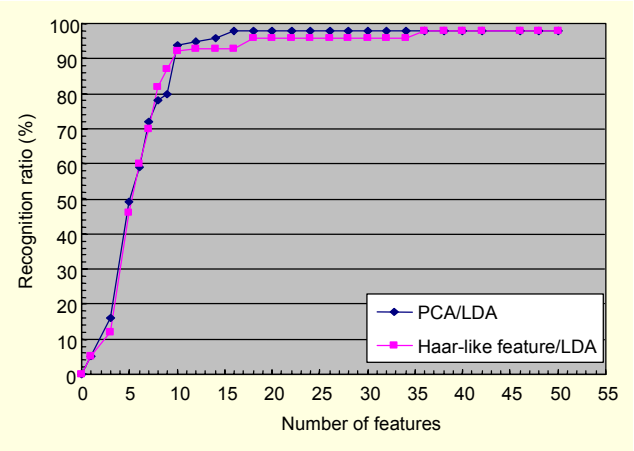

Fig. 6. Comparison of recognition rate between PCA/LDA and the proposed method.

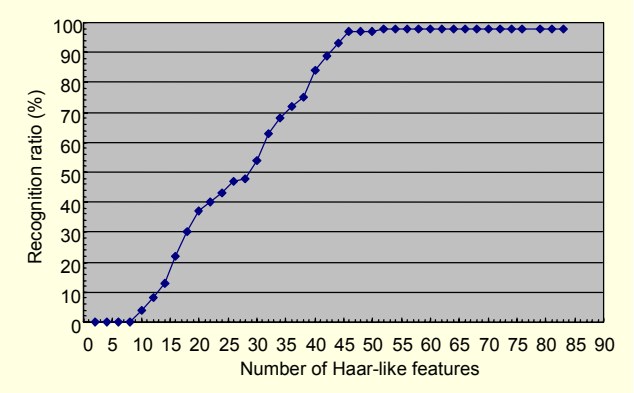

Fig. 7. Recognition rate and the number of Haar-like features.

Table 1. Comparison of training time between PCA/LDA and Haar-like feature/LDA.

	Training time
PCA/LDA	8 min 25 s
Haar-like feature/LDA	37.5 s

Table 2. Processing time of PCA/LDA and Haar-like feature/LDA.

	Processing time
PCA/LDA	1181 ms
Haar-like feature/LDA	93 ms

7 experiments [9]. When comparing the results between Principal Component Analysis (PCA)/LDA and Haar-like feature/LDA, the PCA/LDA recognizer was retrained in the same condition and recalculated the recognition rate in % using rank 1. Figure 6 shows a comparison of the results between the previous method and the proposed one.

As shown in Fig. 6, the recognition rate is increased when the feature number is increased. The recognition rate shows a success of 98% in more than 55 feature numbers. This result proves that the Haar-like feature/LDA can be employed as face recognition, and the recognition rate is similar with the results of the conventional PCA/LDA.

Figure 7 shows the relationship of Haar-like feature numbers and recognition rate when the LDA feature number is fixed as 50. As shown in the results, more than 50 Haar-like features can generate a reasonable face recognition rate.

Table 1 addresses the compared results of training times between PCA/LDA and the proposed method. The number of the feature was fixed as 83 in both experiments to match with the prior experimental results of PCA/LDA. The training time of the proposed method was much faster than that of PCA/LDA because the Haar-like feature can be estimated using a simple summation, while the basis of PCA/LDA requires the computation of a large matrix

Table 2 presents the feature extraction time in both methods for 100 test images. It shows that the feature extraction time of the proposed method is much faster than PCA/LDA.

IV. Conclusion

The ultimate object of this research is dynamic face detection and recognition in real video data. For this purpose, the automatic video management tool was implemented. The resulting tool enables users to access a video easily through generated face information. In addition, the summarization index also supports other operations that can be helpful and interesting for users; querying a scene and editing a video stream. The MPEG-7 descriptors are also used to obtain video summarization and to retrieve a queried scene. The implemented tool was devised to be operated inexpensively with a simple interface; it can be embedded in home electronics such as a personal video recorder. Furthermore, the system can be extended for a video search engine using the MPEG-7 techniques.

References

[1] MPEG-7 Group, *MPEG-7 Applications Document,* ISO/IEC JTC1/SC29/WG11//N2462, Atlantic, Oct.1998.

[2] A. Yamada, E, Kasutani, M. Ohta, K. Ochiai, and H. Matoba, "Visual Program Navigation System Based on Spatial Distribution of Color," *Proc. ICCE,* 2000, pp. 280-281.

[3] Kyeong-Ok Kang, Jae-Gon Kim, Heekyung Lee et al., "Metadata Broadcasting for Personalized Service: a Practical Solution," *ETRI J.*, vol.26, no.5, Oct. 2004, pp.452-466.

[4] T. Walker, "Content-Based Navigation of Television Programs Using MPEG-7 Description Schemes," *Proc. ICCE*, 2000, pp. 272-273.

[5] N. Dimitrova, A. Janeveski, D. Li, and J. Zimmerman, "Who's That Actor? The InfoSip TV Agent," *Proc. the 2003 ACM SIGMM Workshop on Experiential Telepresence*, 2003, pp. 76-79.

[6] T. Meiers, T. Sikora, and I. Keller, "Hierarchical Image Database Browsing Environment with Embedded Relevance Feedback," *Proc. ICIP*, Rochester, NJ, vol. 2, 2002, pp. 593-596.

[7] A. Divakaran, R Regunathan, and K.A.Peker, "Video Summarization Using Descriptors of Motion Activity: A Motion Activity Based Approach to Key-Frame Extraction from Video Shots," *J. Electronic Imaging*, vol. 10, no. 4, 2001, pp. 909-916.

[8] B.H. Jung, M.H. Ha, H.J. Kim, K.S. Park, H.J. Lee, and W.Y. Kim, "A Component-Based DCT/LDA Face Recognition Method for Character Retrieval in TV Programs," *Proc. 5th Int'l Workshop on Image Analysis for Multimedia Interactive Service,* Apr. 2004.

[9] B. S. Manjunath, P. Salembier, T. Sikora, eds, *Introduction to MPEG-7:Multimedia Content Description Lenguage*, John Wiley & Sons Ltd., West Sussex, England, 2002.

[10] M. Yeung and B.L. Yeo, "Segmentation of Video by Clustering and Graph Analysis," *Computer Vision and Image Understanding J.*, vol. 71, no.1, 1998, pp. 97-109.

[11] P. Viola and M. J. Jones, "Robust Real-Time Object Detection," *Technical Report Series,* Compaq Cambridge Research Laboratory, 2001, CRL-2001-1.

[12] T. Cootes, "Model-Based Methods in Analysis of Biomedical Images," *Image Processing and Analysis: A Practical Approach*, R. Baldock and J. Graham, eds, Oxford University Press, 2000, pp. 223-248.