AN INDEXING AND BROWSING SYSTEM FOR HOME VIDEO

*Wei-Ying Ma and HongJiang Zhang**
Hewlett-Packard Laboratories
1501 Page Mill Road, Palo Alto, CA 94304-1126

ABSTRACT

In this paper, we present a content-based video indexing and browsing system for home video. In addition to traditional video parsing to extract temporary structure and abstraction of a video, we also utilize face detection, tracking and recognition techniques to create an organization based on the existence of people in a video. The entire parsing process is designed in the way that it takes full advantage of the information in MPEG compressed data so that it can be performed very efficiently. The system also provides semi-automatic tools for semantic annotation and fully automatic content-based retrieval.

1 INTRODUCTION

In the last few years, many research efforts have been devoted to automatic video content-based indexing, mostly using low-level features and structural information [8]. However, due to the limitation of the low-level visual features in representing semantic content of video, the performance of most proposed content-based retrieval algorithms are far from satisfactory. Recent works in video parsing provide a foundation for building interactive video browsing tools, which complement low-level feature based indexing and searching tools. In the video parsing process, temporal structure in terms of shots is extracted. Browsing can be made more efficient if shots and scenes are represented visually in some abstracted forms, such as key-frames and visual highlights. The limitation of such browsing tools is that it is mostly based on temporal structure, thus does not easily support semantic object based search. Also, most of these video indexing, search and browsing tools require a high degree of user sophistication and, therefore, are suitable only for professional application. There are still no easy tools for home users to organize their home videos and retrieve clips of interest when needed.

In this paper, we present an indexing and browsing system for home video. Home video has many unique features compared to other types of video such as movies, news, and documentary video. Its temporary structure is usually simpler and easier to detect and its content is mostly about people's lives. Therefore, in addition to traditional video parsing to extract temporary structure and abstraction, we also develop algorithms to detect, track, and recognize faces in a video. The system is able to organize home video based on the presence of people. The system also provides semi-automatic tools for semantic annotation and fully automatic content-based retrieval.

2 VIDEO STRUCTURE PARSING AND KEYFRAME SELECTION

Extracting temporal structure of home video in the proposed scheme is the task of video segmentation, that involves the detection of temporal boundaries between scenes and between shots. We have developed a very efficient and robust shot boundary detection scheme that operates directly on MPEG compressed data with very high accuracy and faster than real-time processing speed [4]. The algorithm first looks for a potential shot boundary within a group of pictures (GOP) using the DCT information. This is conducted by comparing the difference between image features extracted from the two consecutive I-frames. The features used in our algorithm include color histogram and average DCT coefficients. If the difference is larger than a certain threshold, a further examination of the exact shot boundary is then conducted using the statistics of macro-block modes in those B- and P-frames in-between. We form $P_{ssb}(i)$ that represents the probability of shot boundary occurred at each frame *i.* For a I-frame, this probability is computed based on the percentage of macro-blocks (MB) with backward prediction mode in its preceding B-frames. For a P-frame, this probability is computed based on the percentage of its own MB with forward prediction mode and the percentage of MB with backward prediction mode in its preceding B-frames. For a B-frame, this probability is computed based on its own MB with forward prediction mode. The frame with maximun probability is then identified and its value is compared with the second largest probability in $P_{ssb}(i)$ and a pre-

* Current address of HongJiang Zhang is: Microsoft Research China, 5F, Bejing Sigma Center, 49 Zhichun Rd. Hadian District, Bejing 100080, P.R.C.

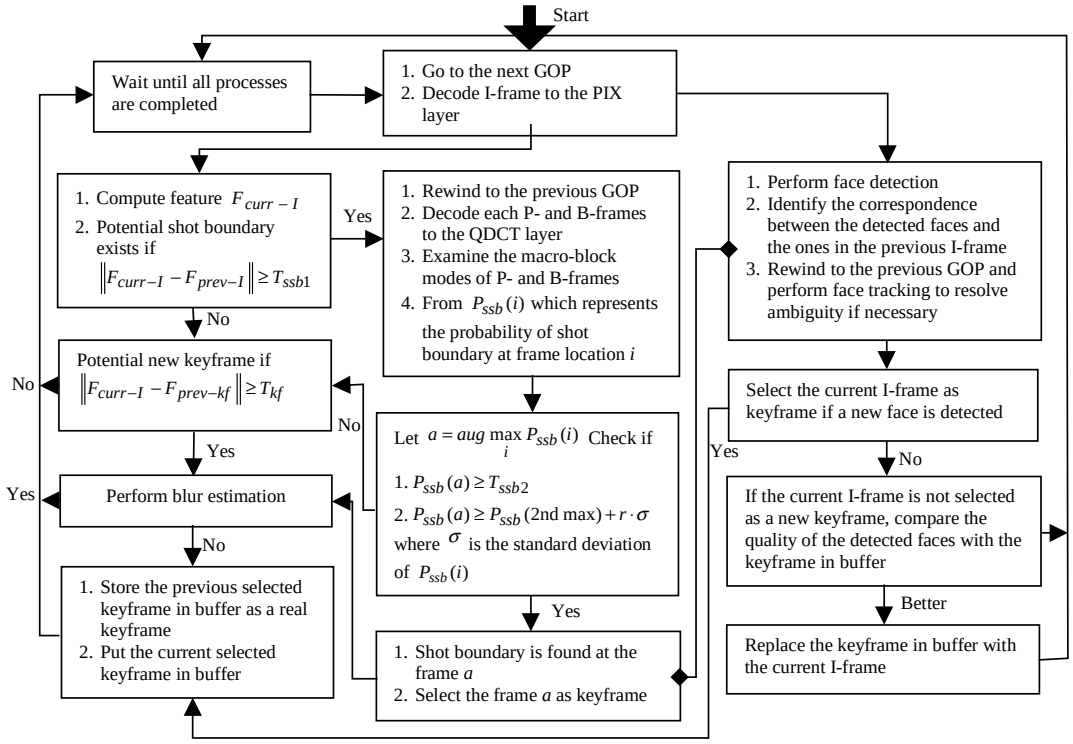

Figure 1. The home video parser efficiently performs shot boundary detection, key-frame selection, face detection, tracking and recognition all together.

defined threshold (see Figure 1). If it satisfies a certain requirement, this frame is marked as a shot boundary. To overcome the problem of selecting thresholds, a statistic framework has been utilized to automatically select the optimal threshold [1][2]. Based on our initial experiments on a Pentium Pro 200 PC with 128MB RAM, the algorithm is able to achieve the processing speed at 160 frames/second with about 95% detection rate.

The key-frame extraction is performed along with the video segmentation process. Basically the algorithm selects the first frame in each shot which passes image blur examination as a key-frame, and continuously looks for the next frame which is significantly different enough from the previous key-frame as a potential new key-frame. Every potential key-frame has to pass through image blur examination [5] before it can be put into the key-frame buffer. This buffer is used to allow a later frame that is similar but has better representation to replace the current selected key-frame. We will discuss this in more details in the following section. Because the blur determination algorithm also utilizes DCT information, it can be performed very efficiently on the MPEG compressed video. By running both shot boundary detection and key-frame extraction, the algorithm is able to achieve a processing speed at 100 frames/second.

The entire process of video parsing is illustrated in Figure 1. It consists of video segmentation and key-frame selection and also face detection, tracking, and recognition that we will discuss in the following section.

3 FACE DETECTION, TRACKING, AND RECOGNITION

People are the most important objects in home video. Therefore, the capability to detect the presence of people, recognize their identities and track them across frames and shots are essential. We have incorporated a neural network-based face detection algorithm developed by Rowley [6] into our video parser. Basically, the face detection algorithm is applied on every I-frame and the locations of the detected faces are matched with the face locations currently under tracking. If a match is identified, the new location of a tracked face is updated and the detected face is labeled as the same person as the tracked face. Whenever there is an uncertainty to identify a newly detected face or there is no match for a current tracked face, the face

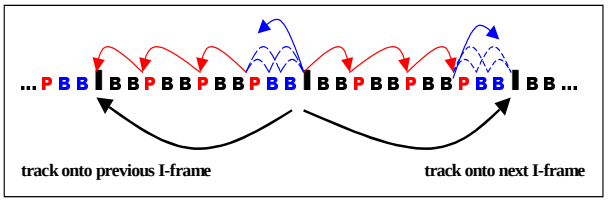

Figure 2. Forward and backward face tracking in MPEG compressed video.

tracking algorithm is applied. We first rewind to the previous GOP and decode all the P-frames and the B-frames after the last P-frame. Then, the motion information from the macro-blocks of these B- and P-frames are used to predict the new location of a tracked face from the previous I-frame to the current I-frame. Figure 2 shows the basic idea of performing face tracking in the MPEG compressed domain. In order to deal with zoom-in and zoom-out, a face region is split into 2x2=4 sub-regions, each tracking one of the four coordinates of the face region. Note that the face detection algorithm can only detect frontal faces, and therefore, it can only indicate the presence of people in many scattered I-frames. The face tracking is essential to identify the correspondence between these detected faces and associate them to a same person within a continuous video shot. The face tracking projects the possible face location from I-frame to I-frame when a person turns his/her face to different angle so that the face detection algorithm fails to detect it. The tracking of a person stops when the projected face region leaves out of scene.

Because the face detection and tracking algorithm can only associate continuous frames that contain a particular person, we need to use face recognition [7] to group disjoined video segments where the same person may appear. Benefited from the rich representation of the faces that we can detect for the same person under different lighting condition and facial expression, the accuracy of recognition can be dramatically improved.

On the other hand, a person may have appeared in a video sequence for some time before his/her frontal face is detected. In order to identify full video segment where this person exists, we also need to conduct backward tracking. To save time, this process is performed after video has been parsed to the end. It then rewinds to the starting location of each detected person and performs backward tracking from there. Frames in a video sequence that contain same (similar) face are grouped automatically, which form the basis for face-based search of home video segments. This allows users to quickly retrieve video based on the query of a particular person or the co-existence of many persons.

Face detection is also used as a condition for key-frame extraction: when the system detects a person entering into a sense, then there could be a need to extract a new key-frame to represent the content. Also, for a set of similar frames, face information is used to help select the best key-frame. That is, a frame that contains larger and better-positioned frontal faces with more people appearing inside can replace the previously selected key-frame if their contents are similar. Face information dramatically improves the performance of key-frame selection algorithm that is based only on the low-level features. Figure 3 shows the result of our video parser using a video sequence of family birthday party. By performing all the previously described tasks together, including video segmentation, face detection and tracking, and key-frame extraction, our algorithm has been able to achieve processing speed at about 10 frames/second.

4 ANNOTATION, BROWSING, AND RETRIEVAL

The previously described face detection, tracking, and recognition can be used to associate the existence of a particular person with a set of video segments. However, in order to provide users a more meaningful way of indexing home video, we need to associate an identity (a name's tag) with that person. We have developed an interactive user interface to facilitate this annotation process [3]. Each time when a video sequence is parsed, the face feature of each detected person is compared against the database and the best match is identified. This match is presented to users for validation. If it is not a correct match, then users will need to type in the correct identity. Otherwise, users can simply click on a confirmation button to complete the annotation process. In the case of misclassification of an existing person, the new face feature is incorporated into his/her face representation to improve future recognition performance. If misclassification is due to a non-existing person, the new name's tag along with the corresponding face feature is included into the database and is used to classify future faces. Since the system is capable of learning different appearance of a person's face under various conditions, such as different facial expression and lighting, the recognition performance can be improved dramatically over time. For home video where the number of people to be classified is usually small (less than 50), our system has achieved a very satisfactory recognition performance.

In addition to the semi-automatic face annotation, the system also allows a user to annotate video based on time and location information in terms of when and

Figure 3. A video browsing system that integrates human face detection, tracking, and recognition

where the video is recorded. The integration of image and text information provides a user a powerful tool for indexing and browsing home video.

5 CONCLUSION

In this paper, we have presented a system for indexing and browsing home video. The system is capable of extracting both structure information and semantic objects such as faces from video. Except that face detection is conducted in image domain, all the other operations such as shot boundary detection, face tracking, and key-frame selection are performed in the compressed domain. Because the system is efficiently and effectively designed to utilize the MPEG compressed data, it is able to process video in almost real time. In addition, we have developed an interactive user interface to facilitate the integration of automated and human annotation, and it results in a hybrid system where computer and users work cooperatively to achieve the best browsing and retrieval performance.

ACKNOWLEDGEMENT

We thank Mike Creech and Alan Kuchinsky for their efforts to incorporate our video processing algorithms into the HP FotoFile system.

REFERENCES

[1] A. Hanjalic and H. J. Zhang, “Shot Detection based on Statistical Optimization.” Tech Report, HP Labs, Sep. 98.

[2] A. Hanjalic and H. J. Zhang, “An Integrated Scheme for Automated Video Abstraction using Unsupervised Clustering.” Technical Report, HP Labs, Sep. 1998.

[3] A. Kuchinsky, et al, “FotoFile: a consumer multimedia organization and retrieval system,” CHI99.

[4] W. Y. Ma and H. J. Zhang, “A video segmentation system using MPEG video data”, Technical Report, HP Labs, Nov. 98.

[5] X. Marichal, W. Y. Ma and H.J. Zhang, “Blur determination in the compressed domain using DCT information,” Proc. of 6th IEEE International Conference on Image Processing, vol. 2, pp. 386-390, Oct. 1999.

[6] H. A. Rowley, S. Baluja and T. Kanade, “Neural network-based face detection,” IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23-38, Jan. 1998.

[7] M. Turk and A. Pentland, “Eigenfaces for recognition,” Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86, 1991.

[8] H. J. Zhang, et al, “Video parsing, retrieval, and browsing: an integrated and content-based solution,” Proc. of ACM Multimedia'95, pp. 15-24, Nov. 1995.