Learning-Based Interactive Video Retrieval System

Chi-Jiunn Wu[1] Hui-Chi Zeng[1] Szu-Hao Huang[1] Shang-Hong Lai[1] Wen-Hao Wang[2]

[1]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
{Legler, mr944327, Howard, Lai}@cs.nthu.edu.tw
[2]ITRI, Hsinchu, Taiwan
devin@itri.org.tw

ABSTRACT

This paper presents an interactive video event retrieval system based on improved adaboost learning. This system consists of three main steps. Firstly, a long video sequence is partitioned into several video clips by using a distribution-based approach instead of detecting shot transition boundaries. Secondly, audiovisual features (i.e., color, motion and audio features) are extracted from video sequences for video clip representation. Finally, the modified AdaBoost learning algorithm is employed for interactive video retrieval with relevance feedback. This AdaBoost learning algorithm differs from conventional AdaBoost learning methods mainly in the selection of paired video features for the weak classifiers. Experimental results show improved performance of video retrieval by using the proposed system.

1. INTRODUCTION

The demand for intelligent processing and analysis of multimedia information has been rapidly growing in recent years. Researchers have actively developed different approaches for intelligent video management, including shot transition detection, key frame extraction, video summarization and video retrieval, etc. Among them, content-based video retrieval is the most challenging and important problem of practical value. It can help users to retrieve desired video segments from a large video database efficiently based on the video contents through user interactions.

The video retrieval system can be roughly divided into two major components: a module for extracting representative features from video segments and one defining an appropriate similarity model to find similar video clips from video database. Many approaches used different kinds of features to represent a video sequence, including color histogram [1], shape information [2], motion activity [3], and text analysis [9]. Some approaches combined some of the above features to improve the retrieval performance [8].

The graph cut technique has been popularly applied for image segmentation in recent years [5]. Some researches [4] extended this idea for video clip segmentation. Another approach used the longest common sequence (LGC) [15] to search for the most similar video sequence.

The relevance feedback from users has been employed to refine the performance in an interactive way. Several efficient learning algorithms have been employed for this purpose in the past, including the AdaBoost algorithm [6] which was developed for the two-class classification problem. The main feature of the AdaBoost algorithm is the capability of selecting the best discriminating features incrementally and adaptively. For the purpose of achieving fast and adaptive online learning for each retrieval task, the AdaBoost learning algorithm is suitable for this kind of problems since the training dataset for each retrieval step is usually very small.

In this paper, a learning-based video retrieval system is developed and described in Sec. 2. This video retrieval system combines color histogram, motion activity and audio features based on an improved AdaBoost learning algorithm with online adaptive paired feature selection for each retrieval task. In sec. 3, the improved performance of the proposed method is demonstrated through experiments. Finally, a brief conclusion is given in Sec. 4.

2. THE PROPOSED VIDEO RETRIEVAL SYSTEM

The proposed video retrieval system is based on online learning the best audiovisual video features for each retrieval task in an AdaBoost framework. In the proposed system, the first step is to partition a long video sequence into several small video clips, i.e. segments where each clip is the basic unit for video retrieval. Secondly, our system extracts three different kinds of video features, including color, motion and audio features based on each clip. Hence, the modified AdaBoost learning algorithm is applied to retrieve the relevant clips from the video database.

2.1. Video Clip Segmentation

In recent years, most video retrieval or scene change detection systems use shots as the basic element in the video database. However, in some cases, a shot could last for a long time with a lot of different activities or subjects, thus it

1-4244-0367-7/06/$20.00 ©2006 IEEE

may not be appropriate to use a shot as the basic unit for video retrieval. For example, a shot in a surveillance video could last for days or hours and there could be a number of different subjects and activities in this shot. Therefore we propose an algorithm that can divide a long video sequence into small clips with each clips corresponding to a period of more uniform activity. In each shot, the distance between the two consecutive frames should be small. Hence, using the distance between every two consecutive frames cannot provide a satisfactory solution to partition a shot further into clips. Thus, each frame needs to compute the distance not only the consecutive frames but also the other frames within a reasonable period in the temporal domain.

We assume each video clip contain at least a minimum number of N frames because very short clips are normally not significant for video retrieval. We first compute a distribution model from the first N frames of the video as the initial distribution for the video clip. The following are the equations for determining the initial clip distribution:

$$D_{ij} = \left| f_i - f_j \right| \tag{1}$$

$$C_\mu = \frac{1}{L} \sum_i \sum_{j=i-W}^{i+W} D_{ij} \tag{2}$$

$$C_\sigma = \frac{1}{L} \sum_i \sum_{j=i-W}^{i+W} \left(D_{ij} - C_\mu \right)^2 \tag{3}$$

where D_{ij} stands for the distance between the i^th^ frame and the j^th^ frame, f_i is the feature vector of the i^th^ frame, C_μ stands for the mean of distance containing the L variables, W denotes a feature distance between the i^th^ frame and the j^th^ frame, and C_σ is the variance of distance containing the L variables.

After modeling the initial distribution, we need to determine if the next frame belongs to this clip or not. Each frame has its own distance distribution. If the frame belongs to this clip, its frame distance will fall within a bound determined by the initial distribution else it will fall outside this bound. The following are the equations to determine the status of the current frame:

$$F_\mu(i) = \frac{1}{2W} \sum_{j=i-W}^{i+W} D_{ij} \tag{4}$$

$$F_\sigma(i) = \frac{1}{2W} \sum_{j=i-W}^{i+W} \left(D_{ij} - F_\mu(i) \right)^2 \tag{5}$$

$$ClipFilter(i) = \begin{cases} Y, & if \left| F_\mu(i) - C_\mu \right| < 2 \times C_\sigma \\ N, & otherwise \end{cases} \tag{6}$$

where $F_\mu(i)$ stands for the mean of distribution in the i^th^ frame and $F_\sigma(i)$ is the variance of distribution in the i^th^ frame. The function $ClipFilter(i)$ is to determine if the i^th^ frame belongs to this clip or not.

If the current frame belongs to the initial distribution, it needs to update the mean and the variance of the distance distribution. Otherwise, the current frame does not belong to this clip, and it needs to compute the next clip initial distribution by equations (1)-(3). Thus we can divide the long video sequence into several clips containing a few main subjects and activities.

2.2. Feature Extraction

A video clip is constructed by a sequence of video frames. The color information is a popular feature in several image and video retrieval system. The color histogram is built by concatenating the histograms of all color channels, i.e. R, G, and B color channels, and each component histogram is sampled into n bins, thus there are $3 \times n$ bins in this histogram. In our system, the color histogram for each video clip is computed by averaging all the color histograms in the same clip.

The other important component of video feature is the motion information. To date, several approaches have been proposed to estimate the motion field. In this system, we apply the diamond search method [12] for motion estimation for its efficiency. This method has been widely employed by many video compression systems. In addition, it is also important to model the motion activity in each frame. Our system considers the motion magnitude and the motion direction to model the motion activity.

For the motion magnitude feature, we compute the motion magnitude histogram to represent the degree of motion at this frame. In human perception, we can easily differentiate static object from moving object with small motions. However, if there are two objects moving with large motion, it is not easy to tell the differences between these two objects. Therefore, we apply the log function to quantize the motion magnitude as our motion features.

In addition, we quantize the motion direction into 8 bins. It is a simple idea to assign each motion direction to the closet reference direction, but the motion direction estimate is not accurate especially when the motion vector is small. This problem will be getting worse when the block-based motion estimation is employed here. To alleviate this problem, we include the motion magnitude information as the weighting factor to compute the weighted motion direction histogram.

As to the audio information, the short-time features of the energy and the average zero-crossing rate [14] have been proved to be effective in discriminating music, speech and silence audio signals. In each video clip, if we only consider the audio information in the key frame, it is not reasonable to describe the content of the audio information. Therefore, we compute maximum, minimum, mean, variance, and histogram of the short-time energy function and short-time average zero-crossing rate of the audio signal in the clip as the associated audio feature vector.

Totally, this system uses 54 features, including 24 color features, 16 motion information and 14 audio features.

2.3. Modified AdaBoost learning algorithm

The conventional AdaBoost learning algorithm [7] [10] combined several simple binary weak classifiers to discriminate relevant and irrelevant datasets. Later, some researchers proposed different ways to modify and improve this learning algorithm to fit to their applications [11] [13]. We employ the improved AdaBoost algorithm [11] in this paper due to its excellent performance in image retrieval. We give a brief description of the main features of this algorithm in the following.

2.3.1. ID3-like balance tree quantization

Previous researches on the AdaBoost learning usually do not focus on feature quantization for density estimation. The main idea of using the ID3-like quantization is to find the best boundary to discriminate relevant and irrelevant samples based on the training dataset. The details of the ID3-like quantization is referred to [11].

2.3.2. Paired feature representation

A unique characteristic of this learning algorithm is the rich information contained in a limited set of features. When the video database contains thousands or millions of video sequence, real-time processing of video features for retrieval becomes a problem. Thus, we propose to adaptively select a small number of paired combinations of effective audiovisual features for the weak classifiers in the AdaBoost learning algorithm. In the traditional AdaBoost algorithm, it employs a single feature for each weak classifier for the two-class classification. In many cases, the distributions of the relevant and irrelevant datasets overlap significantly and can not be separated easily. Using paired feature combination can help to better separate the relevant and irrelevant data samples in a selected 2D feature space, thus leading to a more powerful weak classifier and finally a strong AdaBoost classifier [11].

2.3.3. Bayesian weak classifiers

The original AdaBoost algorithm makes a hard decision in each weak classifier and it may not be very accurate for this binary decision. This modified AdaBoost algorithm [11] applies the Bayesian decision rule to compute the conditional probability as the output of each weak classifier to overcome this problem.

3. EXPERIMENTAL EVALUATION

Our video database contains several types of video, including soap opera, series, NBA games and baseball games. There are totally about eight hours of video for our experiment and we manually label eight events from this video dataset for performance evaluation of video retrieval systems. The eight labeled events are “Coffee shop”, “Home”, “Slam Dunk”, “Haste of basketball game”, “Sound of Cheer”, “Kitchen”, “Living room” and “The action of Pitcher.” In the following, we show how the modified AdaBoost algorithm can provide superior performance on this video retrieval experiment.

3.1. Comparison with different boosting algorithms

Fig. 1 shows the relationship between the average precision of video retrieval and the total number of relevance feedback iterations for the experiment setup described above. Our experiments compare the modified AdaBoost algorithm with the RealBoost and the original AdaBoost algorithms. Note that the scope is set to 36 for each relevant feedback. It is obvious that the modified AdaBoost learning system outperforms the other two boosting systems. Fig. 2 shows the retrieval performance of these learning systems on the “Coffee Shop” and “Slam Dunk” events. The experimental results show that our retrieval system provides superior performance than the other two systems.

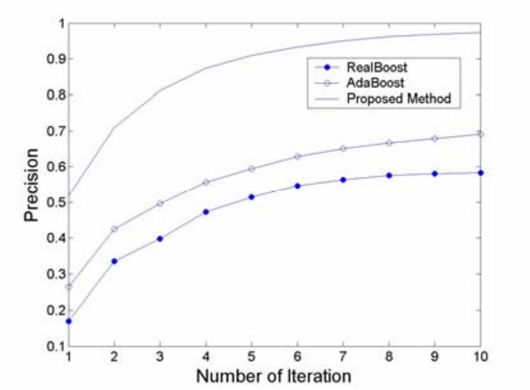

Fig. 1 Precision comparison between different learning approaches from video retrieval experiments with relevant feedback.

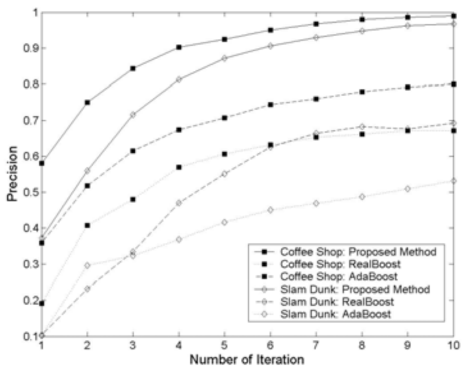

Fig. 2 Precision comparison between different learning methods for video retrieval experiments on the “Coffee Shop” and “Slam Dunk” events.

3.2. Performance of improvement boosting algorithms

We compare the video retrieval performance by using different quantization methods in our video retrieval system with the results shown in Fig. 3. The results show the ID3 quantization method provides superior performance than the other two quantization methods; namely, the histogram equalization method and the uniform quantization method. Fig. 4 shows the performance comparison for using the paired feature and the single feature with different number of selected features in the AdaBoost learning. As shown in the results, the paired feature representation significantly improves the retrieval precision in our experiments.

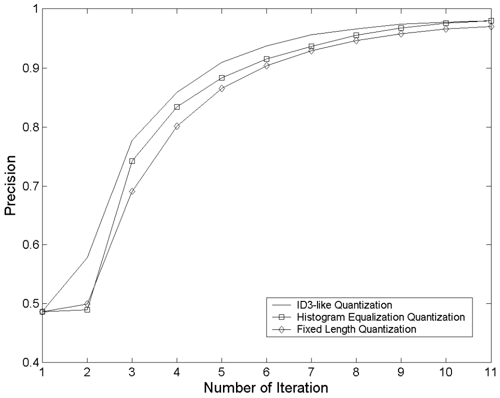

Fig. 3 Video retrieval precision comparison for different quantization methods.

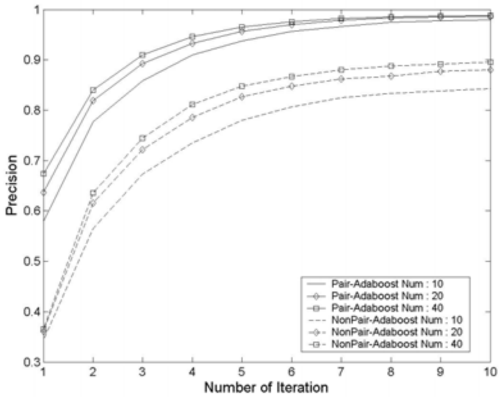

Fig. 4 Retrieval precision comparison for using paired features and single feature in the AdaBoost learning.

4. CONCLUSIONS

In this paper, we proposed a novel video event retrieval system based on on-line AdaBoost learning. We first partition the video data into video clips and extract the representative audiovisual features for each video clip. Then we employed the improved AdaBoost learning algorithm to iteratively refine the retrieval results via relevant feedback. Our experimental results show that the proposed video retrieval system can provide accurate retrieval results and the modified AdaBoost learning significantly improves the video retrieval performance. The modified components in our retrieval system, including the ID3 quantization and paired feature combination, are verified to contribute performance boosting considerably.

ACKNOWLEDGEMENTS

The authors wound like to thank ITRI EOL's support for this research.

REFERENCES

[1] A. M. Ferman, A. M. Tekalp, and R. Mehrotra, "Robust color histogram descriptors for video segment retrieval and identification," IEEE. Trans. on Image Processing, Vol. 11, No. 5, pp 497-508, 2002.

[2] B. Erol, and F. Kossentini, "Shape-based retrieval of video objects," IEEE, Trans. on Multimedia, Vol. 7, No. 1, pp 179-182, 2005.

[3] C.W. Ngo, T.C. Pong, H.J. Zhang, "Motion-based video representation for scene change detection," Int. Journal Computer Vision, pp 127-142, 2002.

[4] C.W. Ngo, Y.F. Ma, and H.J. Zhang, "Video summarization and scene detection by graph modeling," IEEE T. Circuit System Video Tech., 2005.

[5] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral Grouping Using the Nystorm Method," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, pp 214-225, 2004.

[6] Y. Freund, and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Computational Learning Theory: Eurocolt, pp 23-37, 1995.

[7] K. Tieu and P. Viola, "Boosting image retrieval," Proc. IEEE Conf. Computer Vision Pattern Recog., pp 228-235, 2000.

[8] L. Chen and T.S. Chua, "A match and tiling approach to content-based video retrieval," Proc. ICME, pp. 301-304, 2001.

[9] M.Y. Chen and A. Hauptmann, "Searching for a specified person in broadcast news video," Proc. ICASSP, Vol. 3, pp 1036-1039, 2004.

[10] P. Viola and M. Jones, "Rapid object detection using boosted cascade of simple features," Proc. IEEE Conf. Computer Vision Pattern Recog., pp 511-518, 2001.

[11] S.-H. Huang, Q.-J. Wu, and S.-H. Lai, "Improved Adaboost-based image retrieval with relevance feedback via paired feature learning," Intern. Conf. on Image and Video Retrieval, pp 660-670, 2005.

[12] S. Zhu and K.K. Ma, "A new diamond search algorithm for fast block-matching motion estimation," IEEE Trans. Image Processing, Vol. 9, pp 287-290, 2000.

[13] S. Z. Li and Z. Zhang, "FloatBoost learning and statistical face detection," IEEE Trans. Pattern Analysis Machine Intelligence, Vol. 26, pp 1112-1123, 2004.

[14] T. Zhang and C.-C. J. Kuo, "Audio content analysis online audiovisual data segmentation and classification," IEEE Trans. Speech and Audio Processing, Vol. 9, No. 4, pp 441-457, 2001.

[15] Y. T. Kim, and T. S. Chua, "Retrieval of news video using video sequence matching," Proc. of 11th MMM Conference, pp 66-75, 2005.