Speaking Video Summary Based on Face Detection in Moving Region

Huang Jianqiang*, Wang Xiaoying, Cao Tengfei and Wang Rui

Department of Computer Technology and Application, QingHai University, Xining 810016, Qinghai, China

Abstract: Static video abstract is able to show the dynamic event of semantics. According to the current research of events, the method of generating speaking video summary by detecting face in moving region is proposed. Extracting face-based frame and its moving region, combined with the trained face classifier, the speaking event is detected. Integrated video temporal and spatial characteristics, this method expresses the visual impact of video contents effectively. Experimental results indicate the generated summary with good effects.

Keywords: Face detection, moving region, speaking event, video summary.

1. INTRODUCTION

Digital video, as an important form of information in the future, has been used over television, Internet, mobile phones and other media in large numbers. Although large-capacity, high-performance and high-bandwidth server can alleviate the problem of storing video data, but how to allow users to enjoy the convenience of video analysis service has become the focus of people's attention. However, the content of the video itself not only takes a huge storage space, and at the same time contains rich physical structure contains a complex logical structure [1], which impede the video content analysis.

In recent years, video summaries [2, 3], which is developing rapidly, can reflect most important information in some simple ways form the main content of the video, that has met the needs of users. Events as "the world's space-time occurrence once," [4], it is a concrete manifestation of video semantic understanding, compelling and meaningful highlights in particular video sports, news, feature films, monitoring, etc… In the news video [5, 6], the multicast frame by announcer mouth, face, subtitles and other information segmentation of different events. In contrast, movies, the plot of TV and other feature films ups and downs, and full of affectionate talking, fierce fighting, explosions and other exciting screen, and in the structure but no potential areas of knowledge available. In document [7], the use of subtitles and movie script segmenting some meaningful scene firstly, and then through the important function (including keywords, character names, the protagonist), it can identify important scenes. In document [8], through offline split the movie into a camera shot and calculate its characteristics, and then obtain online and analyze interesting camera shot selected by the user, and producing video summaries contain some semantics.

In document [9], it combines the underlying sound, visual and video caption feature to detect events, produce a video summary from the bottom up.

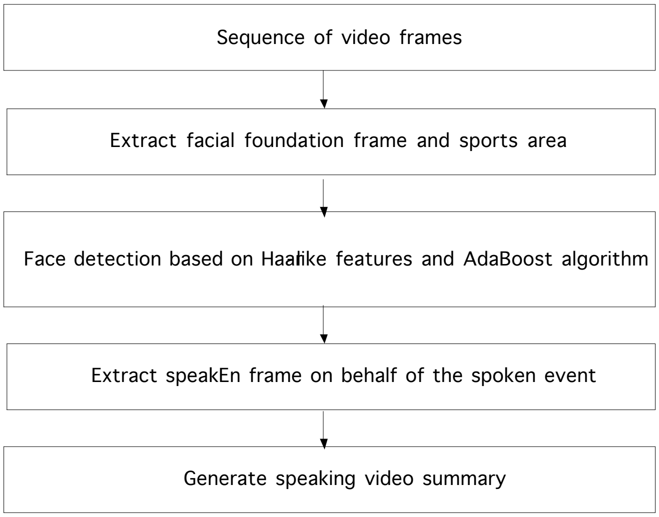

Sequence of video frames
Extract facial foundation frame and sports area
Face detection based on Haarlike features and AdaBoost algorithm
Extract speakEn frame on behalf of the spoken event
Generate speaking video summary

Fig. (1). Process of generating speak video summary.

Meanwhile, the method that generates a video summary through the event detecting depend on specific applications, and it still inseparable from the underlying characteristics of effective convergence and middle semantics. In this paper, we begin from a video character (such as news, feature films, etc.) which contain visual characteristics, and detect talk event by analyzing the existence of a face in the motion area, It's not only to straightly recognize the character quickly, and also can further understand the behavior of and the accompanying emotions and other details. The video summary of the talking human face detection based on motion area, generation process is as shown in Fig. (**1**).

2. EXTRACT THE FUNDAMENTAL FRAME OF HUMAN FACE AND MOTION AREA

Because there are a lot of redundant video sequence frames in the timeline, and the object of interest is usually the movement, therefore, this section uses the "symmetric difference" [10, 11] method which calculation is simple and fast, especially for dynamic scene changes to detect the motion intensity, and filter the majority of similarly frames (video frames retained as "basic face frame"). In order to

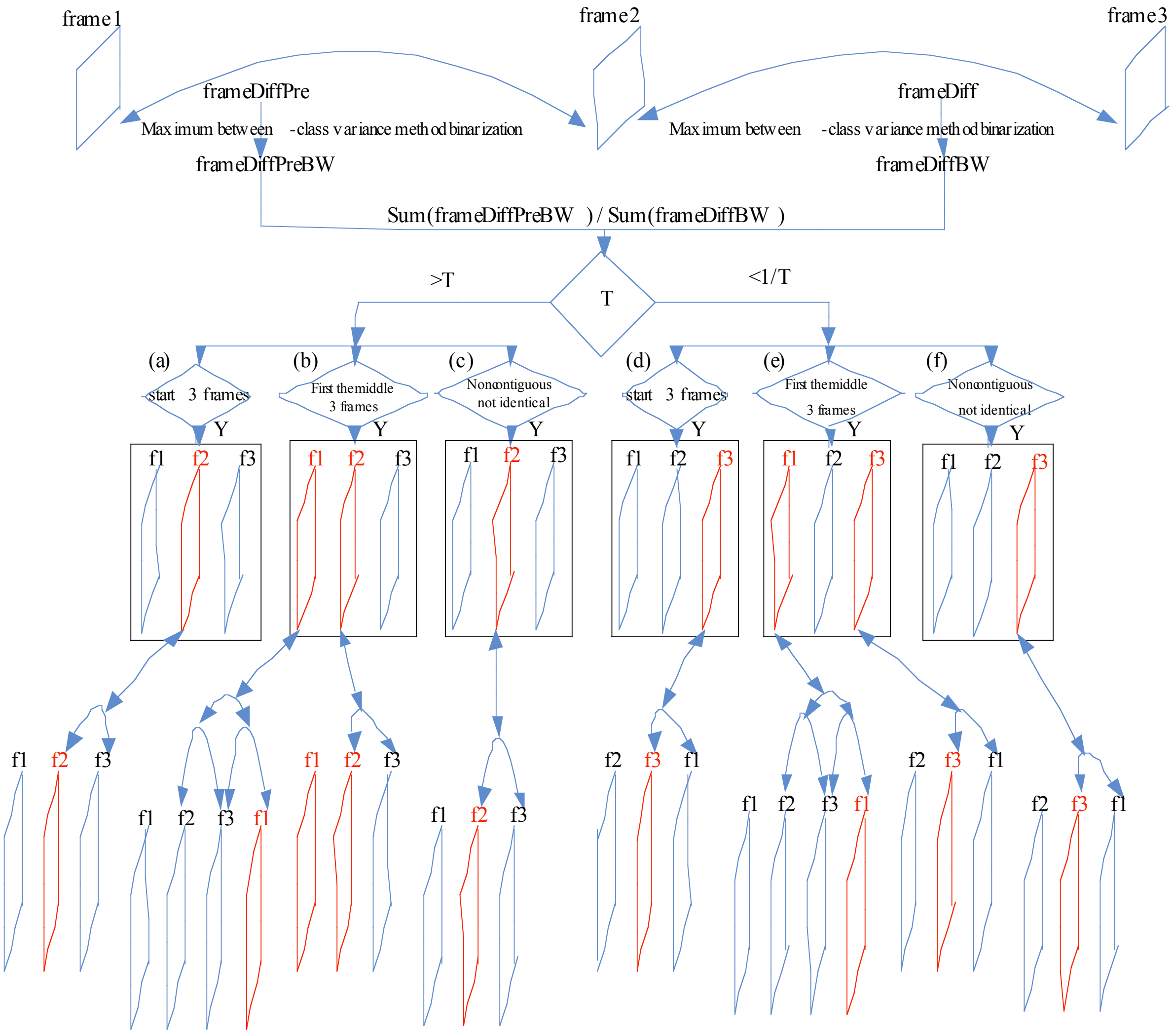

frame1
frame2
frame3
frameDiffPre
frameDiff
Maximum between -class variance method binarization
Maximum between -class variance method binarization
frameDiffPreBW
frameDiffBW
Sum(frameDiffPreBW) / Sum(frameDiffBW)
>T
T
<1/T
(a)
start 3 frames
Y
(b)
First the middle 3 frames
Y
(c)
Noncontiguous not identical
Y
(d)
start 3 frames
Y
(e)
First the middle 3 frames
Y
(f)
Noncontiguous not identical
Y
f1 f2 f3
f1 f2 f3
f1 f2 f3
f1 f2 f3
f1 f2 f3
f1 f2 f3
f1 f2 f3
f1 f2 f3 f1
f1 f2 f3
f1 f2 f3
f2 f3 f1
f1 f2 f3 f1
f2 f3 f1
f2 f3 f1

Fig. (2). Scheme of extracting face foundation frame and motion region.

reduce the effects of random noise on exercise intensity estimation, the differential image "maximum between-class variance" binary processing. Then, judge the fundamental frame of human face according to the change in the binary image of a moving object and change direction.

Motion area is a Part of the fundamental frame of human face and so extract them at the same time, and it need to refer to the adjacent frame difference before and after images, to map the motion area correctly to the corresponding fundamental face frame.Taking into account that the user's attention mainly focused on the center or on both sides the video screen, moving regions are simplified to three default locations. Specific extraction method has shown in Fig. (**2**):

Solution varies with the degree and direction of change of moving objects in video sequences, is divided into greater than T and less than 1 / T in both cases (the greater in the threshold value T, the greater changes of the face frame basis). Fundamental frame of human face marked in red frame, arrows correspond to its extraction strategy moving region. For example in Fig. (**2**) (b) the corresponding branch is firstly greater than T in three consecutive frame images of a video sequence in the middle, indicating a large change extent between f1 and f2 and f3 f2 and extract f1 and f2 as the fundamental frame of human face. Since the left side of f1 hasn’t detected the fundamental frame of human face, then in the left area differences of f2 and f3 and the frame difference of f1 and f3 as the phase difference image frame operation can more accurately extract the f1 corresponding motion area. The motion area of f2 extract corresponding to the right of the image from the frame difference of f2 and f3.

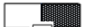
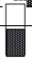
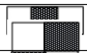
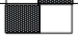
In order to face its basic frame feasibility motion region extraction program verification, select some open-video and youku videos, the experimental results shown in Table **1**:

In Table **1**, the omission frames of the fundamental frame of human face is mainly because small degree of change between frames of a video screen or a change not in the center, generally cannot attract the user's attention. Repeat mainly because almost no change in the adjacent two frames in the third large relative changes caused by sudden. Compared with Storyboard frames and the percentage that open-video provided, the plan retains the full face foundation frame and sports area, while greatly reducing the redundancy, reflecting the visual attention point, has a strong feasibility .

Table. Character-based frame extraction results.

Video Clip Name	The Total Number of Frames	Storyboard Frames	Face Foundation Frames				Screening Percentage
			Observes	Program	Omission	Repeat	
Hovercraft	3814	6(open-video)	23	23	2	2	0.60%
Public Damage	10791	14(open-video)	52	50	7	5	0.46%
Fei Cheng	838	(youku)	8	7	2	1	0.84%

Table 2. Haar-like feature extraction settings.

Characteristics of Prototype	The Original Datum Features a Rectangle			Rectangular Scaled Baseline Characteristics			Characteristic Quantities	Total Characteristics
	Starting Location	h	w	Starting Location	$h^{'}$	$w^{'}$		
	(2, 2)	3	6	(2*q, 2*q)	3*q	6*q	112	358
		6	3		6*q	3*q	112	
		3	9		3*q	9*q	70	
		6	6		6*q	6*q	64	

3. HUMAN DETECTION BASED ON HAAR-LIKE FEATURES AND ADABOOST ALGORITHM

Description of the facial features, skin color [12, 13] quickly and efficiently, but the regional differences between different parts of the face can be divided into more than a single pixel, thus Haar-like [14] describe accurately characterized and real-time unanimously approved on.

This section uses the features four prototypes in the movement area of the face to extract the basic frame rectangle feature detection window by calculating the integral image in document [14].

Wherein, filtered a small rectangular little discriminating ability to reduce the feature dimension, simultaneously, without increasing the amount of additional calculations to change the case size of the detection window to face the multi-scale problem solving.

Specific set of feature extraction as shown in Table **2** (wherein, h, w are the height and width of the rectangle feature, q is the scaling factor):

Table **2** Central baseline characteristics of the rectangle starting position from (2, 2) to start, in order to avoid the effects of facial image non-face portion of the corners. Characterized by the ratio of the zoom rectangle, in order to ensure the same number of feature extraction, detection window is required to start position and characteristics of the prototype together with the same height and width ratio scaling.

In extracted rectangular feature vector, each dimension represents a partial facial features, and a human face is week classifiers. Within these weak classifiers, some judgment as to whether a human face plays a key role, while others are superfluous. Therefore, considering the numerous train a weak classifier by certain upgrade strategy generated a strong classifier, human face detection services. This section uses based on the "Two hands are better than one" thought the AdaBoost algorithm, MIT-CBCL Face Database [15] extracted Haar-like feature set of samples and classification label set on a classic face database. If the sample correctly classified by its smaller weights, weights otherwise unchanged, this emphasis on gradual weight update rule misclassified samples, enhancing the relevance and accuracy of the training, the training process in detail see [14].

About setting the threshold weak classifiers method did not discuss the relevant literature, this section will analyze with the nearest neighbor based segmentation criteria and orderly exhaustive set of weak classifiers threshold.

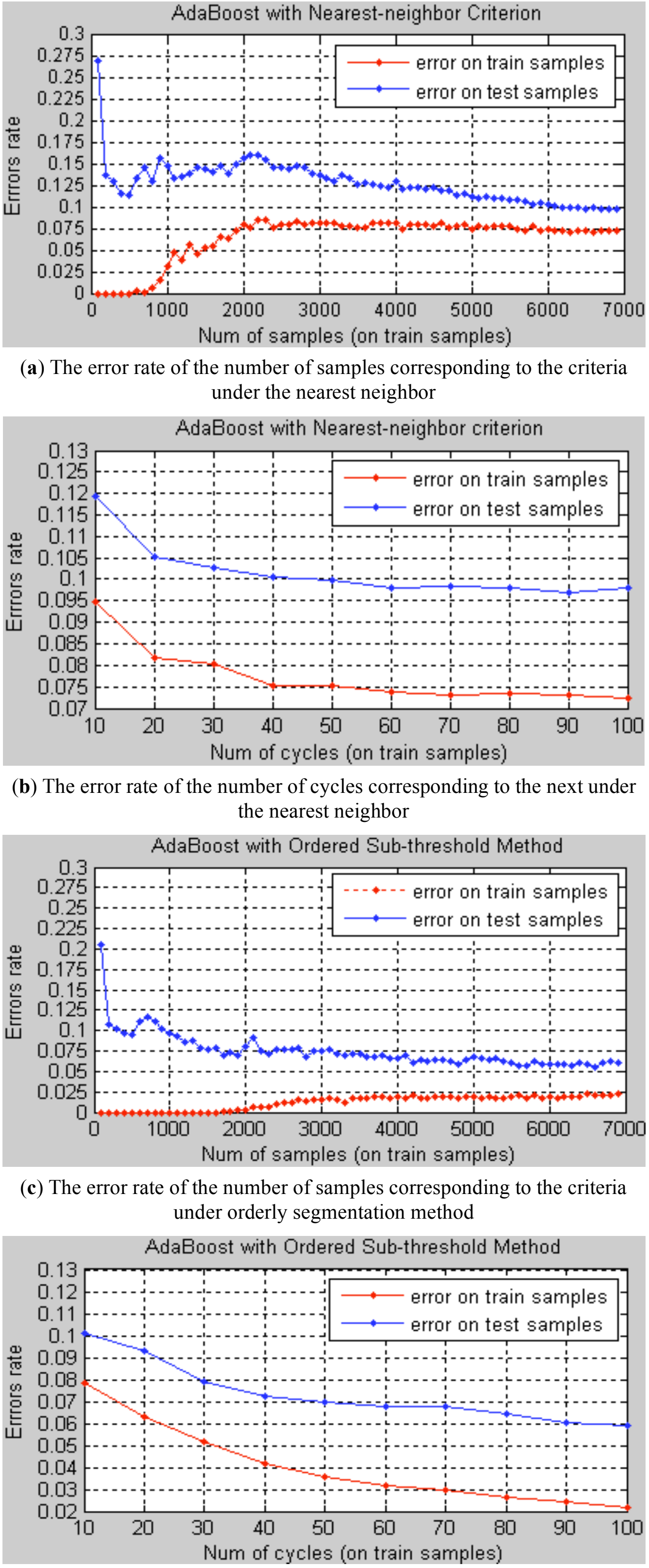
Based on the criteria that set weak nearest neighbor classifier threshold means that if a ample mean of the sample set from the face of more recent human face samples are returned, otherwise classified as non-face samples. Accordingly to this,j-th weak classifier is designed to:

$$\theta_j = \frac{\text{mean}(x_{ij}\,|y_i = 1) + \text{mean}(x_{ij}\,|y_i = 0)}{2}$$

$$h_j = \begin{cases} 1 \;\; \textit{if} \;\; \text{mean}(x_{ij}\,|y_i = 1) > \text{mean}(x_{ij}\,|y_i = 0) \; \textit{and} \; f_j(x) > \theta_j \\ 1 \;\; \textit{if} \;\; \text{mean}(x_{ij}\,|y_i = 1) < \text{mean}(x_{ij}\,|y_i = 0) \; \textit{and} \; f_j(x) < \theta_j \\ 0 \;\; \textit{others} \end{cases} \quad (1)$$

where, x_{ij} is the i-th sample of the j-th feature, y_i and h_j are the classification label and the classification result (1 denotes a face, 0 indicates the non-face), θ_j is the j-th threshold of the weak classifiers, $f_j(x)$ is the j-th characteristic value of a sample of the weak classifiers.

Ordered segmentation based on finite exhaustive set of weak classifiers is the threshold between the minimum and maximum sample set, such as the j-dimensional feature is



(**a**) The error rate of the number of samples corresponding to the criteria under the nearest neighbor

(**b**) The error rate of the number of cycles corresponding to the next under the nearest neighbor

(**c**) The error rate of the number of samples corresponding to the criteria under orderly segmentation method

(**d**) The error rate of the number of cycles corresponding to the next under orderly segmentation method

Fig. (3). Two weak classifiers corresponding threshold setting method detects an error rate.

divided into N segments, each section were exhaustive endpoint value as the threshold of j-th weak classifiers, the expression is:

$$\theta_j = \min(x_i) + j * \frac{\max(x_i) - \min(x_i)}{N}, 1 \le j \le N$$

$$h_j = \begin{cases} 1 & f_j(x) > \theta_j \\ 0 & others \end{cases} \qquad (2)$$

Using these two respectively settings weak classifier threshold approach in training from a number of rectangular features derived from the 100 best combination of weak classifiers to enhance a strong classifier, then, test on the training and test sets of MIT-CBCL Face Database and the detection error rate as shown in Fig. (**3**):

In Fig. (**3**), along with the number of samples or optimal number of weak classifiers (*i.e.*, the number of cycles) continues to increase, the resulting strong classifier performance and improve the detection error rate continues to decrease when it reaches a certain level, detection error rate essentially unchanged. On the set of weak classifiers threshold error rate, the method based on the most recently detected in the neighborhood criterion of training sample set is about 7.25%, about 9.75% on a test sample set, relatively short training time, and ordered based segmentation exhaustive method detects an error rate of 2.25% and 6.05% respectively, the training time is relatively long. Between the number of samples and the presence of the optimal number of weak classifiers on the detection error rate consistency, 100 generated by the optimal combination of weak classifiers has been elevated sufficiently strong classifier determines from the entire face of the sample set. Thus, the selection of a lower error rate detection segmentation based on orderly exhaustive strong classifier trained lay the foundation for detecting the video speak of the event.

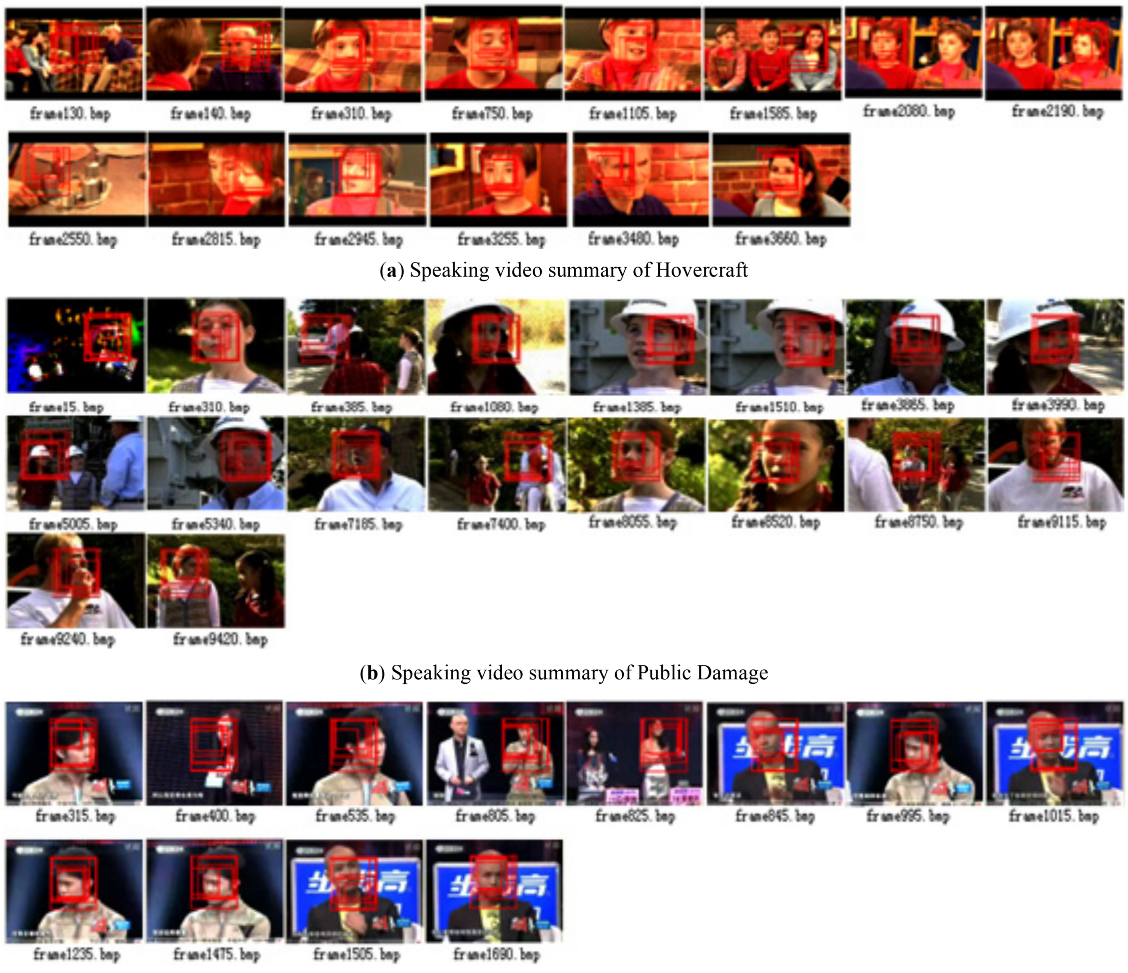
4. THE GENERATION OF SPEAKING VIDEO SUMMARIES

"Speak", which is expressing with language can reflect a person's attitudes, feelings and psychology. In the detection of the video stream, spoken events are as middle semantic category, closer to the user's understanding. This section uses the characters on the basis of a video frame moving area has good face classifier training event detection speaking, if a face is detected, the frame is added as a speaker to speak on behalf of the video summary, generated results are summarized in Fig. (**4**):

In Fig. (**4**), red box speak calibrate the area that speaking event occurs, the quality of spoken video summaries characters in a video stream by frame basis the motion accuracy of the calibration area, face reflection, complex background and other factors, with the limitation of experimental conditions, number of rectangular features that use for training face classifier when using greatly reduced relative to the total number of features there has been missed and the error detection portion. Overall, spoken video summary reflects the character's speaking event in video, shows the video content to the user with a visual shock effect, and has a certain value.

CONCLUSION

For video summary of events can not only express the "what" of a static entity, and also can be further summarized



(**a**) Speaking video summary of Hovercraft

(**b**) Speaking video summary of Public Damage

(**c**) Speaking video summary of "You Are The One"

Fig. (4). Example of spoken video summary.

the dynamic behavior- "What are you doing?" Firstly, propose the method of extracting the character fundamental frame and the possible motion area in a video sequence, then, cut and set reasonable description of the human face Haar-like features, presented and analyzed two weak classifier threshold setting method for AdaBoost Algorithm, at last, use the trained classifier to detect spoken face event in the motion region which based on the fundamental frame of human motion. Experimental results show that this method generated abstract with good results.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their helpful advices to improve this paper.This work was supported in part by a grant from the Applied basic research of Qinghai Natural Foundation (2014-ZJ-718), Youth Foundation Program of Qinghai University (2013-QGY-10), Youth Program of Qinghai Natural Foundation (2013-Z-930Q) and National Natural Science Foundation of China (No. 61363019),Youth Foundation Program of Qinghai University (2014-QGY-23).

REFERENCES

[1] Sun Zhonghua. Reseach on Content Based Video Structure Analysis and Abstraction [D]. Jilin University, 2007. (in chinese)

[2] Money A G, Agius H. Video summarisation: A conceptual framework and survey of the state of the art[J]. Journal of Visual Communication and Image Representation, 2008, 19(2): 121-143.

[3] Truong B T, Venkatesh S. Video abstraction: A systematic review and classification[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), 2007, 3(1): 3.

[4] Jain R. Experiential computing[J]. Communications of the ACM, 2003, 46(7): 48-55.

[5] JIANG Fan,ZHANG Yu-Jin. News Video Indexing with Scene Sectioning and Summary Generation [J]. CHINESE JOURNAL OF COMPU TERS, 2003,26 (7),859-865. (in chinese)

[6] XIE Yu-xiang,LUAN Xi-dao,WU Ling- da,XIAO Peng. A Method of News Video Summarization Based on EDU Model [J]. JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY, 2007,29 (5) :71-76. (in chinese)

[7] Tsoneva T, Barbieri M, Weda H. Automated summarization of narrative video on a semantic level ICSC 2007. International Conference on. IEEE, 2007: 169-176.

[8] Mehdi Ellouze, NozhaBoujemaa, Adel M. Alimi, Interactive movie summarization system, Journal of Visual Communication and Image Representation, 2010, 21(4): 283-294.

[9] G. Evangelopoulos, A. Zlatintsietal, Video event detection and summarization using audiovisual and text saliency, IEEE Interna-

tional Conference on Acoustics-Speech and Signal Processing,2010,3553-3556.

[10] SONG Hong,SHI Feng,WANG Yizhuo. Method of Real-time Face Detection in Video Images [J]. Computer Engineering,2004, 30(19):23-24. (in chinese)

[11] Yin Yong , Liu Xi-fu. Face detection and prediction technology based on video surveillance [J]. Journal of Chongqing University, 2008,31(7):786-791.(in chinese)

[12] HSU R.L,Abdel MottalebM,Jain A K.Face detection in color images[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2002,24:696-706.

[13] CHEN Duan-Sheng,LIU Zheng-Kai. A Survey of Skin Color Detection [J]. CHINESE JOURNAL OF COMPUTERS, 2006,29 (2) :194-207. (in chinese)

[14] PaulViolaRobust Real-Time FaceDetection,International Journal of Computer Vision,2004,57(2):137-154.(in chinese)

[15] MIT-CBCLFaceDatabase.http://cbcl.mit.edu/software-datasets/FaceData2.html

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014