J-007

Analaysis of Human Gaze in Videos with Human Face – Comparison of Gaze Distance with Detected Face and Motion Center

Kok-Meng Ong †　Wataru Kameyama †

1. Introduction

With the exponential increase of digital video content, the field of digital video content analysis has been actively researched. Researchers have been trying to bridge the semantic gap of video with its bit streams using myriads of approaches utilizing the low-level signals [1]. Because the ultimate user of video content is human viewer, the Region Of Interest, ROI, of human towards the content shall not be disregarded in understanding the video context. Therefore, in this paper, we report on a study that compares 2 types of automatically generated ROIs. Actual human gaze is collected and compared with the automatically generated ROI, by finding their distances for videos that contain human faces.

2. Region of Interest

In general, ROI is a portion of a video that viewers show more interest in or pay more attention to than others. A precise definition of an ROI in a video frame is: a spatial portion of a frame that contains the key concept or main subject of a visual scene and provides end users with a more concise and informative representative of a frame, e.g., the speaker should be one of the ROIs in a conference scene [2]. In this paper, two types of ROIs are automatically generated as follow.

2.1 Face

Human viewer usually pays more attention to the human face in video. Therefore, detecting the face as ROI is of practical interest. Haar classifier is used as the face detector in this paper, and the training set provided by OpenCV is used [3]. The locations of possible faces are detected for each frame in the video sequence. Because false detection and failure in side faces exist in this frame-based algorithm, the spatial and temporal property of video is further exploited to improve the face detection performance as follow:

2.1.1 Spatial Constraint

First, the overlapped detection is filtered out by applying spatial constraint.

$$D(C1,C2) > \alpha \times R$$

Where $D(C1,C2)$ is the distance of the center between two detected face location. α is the threshold and R is the average radius of the detection circle. The spatial filter is applied here in order to filter out the false overlapped detection, because practically two faces could not appear together spatially.

2.1.2 Temporal Constraint

Second, detection noise is filtered out by applying temporal constraint of shot.

$$T_{\text{det}} > \beta \times T_{shot}$$

Where T_{det} is the total duration that the detected region appears, T_{shot} is the shot duration and β is threshold. The temporal constraint is employed here based on the definition that video shot is continuous frames that are taken from a camera. Therefore, for a detection to be valid, the time constraint is applied here so that it must appear for more than a certain ratio of time to the shot duration.

2.1.3 Linear Interpolation

Finally, linear interpolation is carried out within a shot to recover undetected faces or profile face.

2.2 Video Motion Center

The second ROI suggested in this paper is the motion center. Our hypothesis here is that when there is object movement in the video content, ROI will follow the center of movement of the video.

The video's center of motion is extract based on the steps below [4]:

2.2.1 Interframe Difference

Calculate the inter-frame difference for each pixel, PD:

$$PD_{x,y|^t} = P_{x,y|^t}(Y,U,V) - P_{x,y|^{t-1}}(Y,U,V)$$

Where $P_{x,y|^t}(Y,U,V)$ is the pixel value at time t.

2.2.2 Moment Calculation

Find the center of motion, (X,Y) by calculating the moment:

$$X = \sum_{x}^{width} \sum_{y}^{height} PD_{x,y} \times x$$

$$Y = \sum_{x}^{width} \sum_{y}^{height} PD_{x,y} \times y$$

† Graduate School of Global Information and Telecommunication Studies, Waseda University

3. Experiment

6 volunteers participated in the experiment. The subjects were shown a comedy-drama film, The Devil Wears Prada, which features more appearance of human face and conversational dialog. The subjects gazing points were recorded from the participants eye using VIS-EYE Measurement System at 60 Hz sampling frequency [5]. The pupil sizes were recorded during the experiment but are not used for analysis in this paper. Refer to our previous work [6] for more information regarding the experimental setup.

3.1 Five Types of Video Scene That Involve Human

The gradation of distances between the camera and recorded subjects can be infinite. However, according to Arijon [7], actual practices has thought that there are five basic definable distance for video shots that feature appearances of human character:

- Close Up, or Big Close Up
- Close Shot
- Medium Shot
- Full Shot, and
- Long Shot

Therefore, five video clips, which include human appearance with the abovementioned distances, are extracted from the movie that was shown to the test subjects, and are analyzed in this paper.

4. Result

For comparison of the two generated ROIs, the Euclidean distance with the actual gazing points of the test subjects are calculated. For each of the 5 types of video, the average distance of all the 6 subjects are tabulated in Table 1. The overall average distance between the gazing points and the center of detected face is 32.14 pixels and with the extracted movie motion center is 52.07 pixels.

5. Discussion and Future Work

From the result, it can be seen that generally the distance of human gaze is closer to faces in video instead of movie motion center. The difference is significantly larger in Close Shot and Medium Shot video type.

For Close Up shot, although the gaze is closer to the recognized face center, the different is relatively smaller. This is because the face itself covered most of the video frame, the viewer tends to look at the eyes or mouth, wheareas our method take the center of the face as the point for comparison. Therefore, a finer localization for Close Up shot could possibly improve the determination of ROI, and further reduce the distance.

Table 1: Comparison of the Distance between Human Gaze and the Generated ROI

Types of Video	Euclidean Distance with Actual Human Gaze In Pixel (Average for all Test Subjects)	
	Face	Motion Center
Close Up	30.30	42.52
Close Shot	26.82	60.65
Medium Shot	27.99	53.47
Full Shot	43.36	51.62
Long Shot	Face Undetected	
Overall Average	32.14	52.07

However, the method used here failed to detect face in Long Shot video type. For Long Shot video type, the motion is relatively higher if compared to to the other video types. After applying filtering methods in Section 2.1, the detections were filtered out. Therefore, further study is needed to verify if the ROI of this video type could be determined with motion center.

6. Conclusion

In this paper, the two types of ROI in video frame are studied, namely the face and video motion center. Subjective evaluation was carried to compare the ROIs. Experimental result shows that human gaze is relatively closer to the face than movie motion center in videos that contain human faces. The finding could be used as an indication for the human interest in future video content analysis studies because when the viewer is attentive to the content, their gaze will naturally tracking the ROI.

Acknowledgement

This research has been conducted jointly with VIS Research Institute.

References

[1] N. Dimitrova, H.J. Zhang, B. Shahraray, I. Sezan, T. Huang and A. Zakhor, "Application of Video-Content Analysis and Retrieval", IEEE Multimedia, Vol. 9, No. 3, pp. 42-55, 2002

[2] Wen-Huang Cheng, Wei-Ta Chu, and Ja-Ling Wu, "A Visual Attention Based Region-of-Interest Determination Framework for Video Sequences", IEICE Trans. Inf. & Syst., Vol. E88-D, No. 7, pp. 1578-1586 July 2005

[3] Gary Bradsky and Adrian Kaehler, "Learning OpenCV: Computer Vision with the OpenCV Library", O'Reilly Media, Sebastopol, September 2008: First Edition.

[4] Doi Shigeki, "Beginning Video Programming", CQ Publisher, 2007, (In Japanese)

[5] Visual Interactive Sensitivity Research Institute http://www.visri/jp/english/index.html

[6] Kok-Meng Ong and Wataru Kameyama, "Classification of Video Shots Based on Human Affect", The Journal of The Institute of Image Information and Television Engineers, Vol. 63, No. 6, pp. 135-144, 2009

[7] Daniel Arijon, "Grammar of the Film Language", Silman-James Press, Los Angeles, 1976