Proceedings of the Fifth IIEEJ International Workshop
on Image Electronics and Visual Computing 2017
Da Nang, Vietnam, February 28 - March 3, 2017

SELECTION OF LOCALIZED AUDIO TRACK BASED ON EYE-TRACKING TECHNOLOGIES WITH APPLICATION TO MUSICAL ART GALLERY

Fumiya Shimizu Issei Fujishiro

Graduate School of Science and Technology, Keio University

ABSTRACT

Many methods using eye tracking have been proposed for realizing interfaces for everyday device including digital signage and HMD. Such visual interfaces have a potential to be utilized as audio interface. We herein intend to allow the viewer to select the audio data by detecting his/her gaze with a single webcam. As a result, we will be able to provide the viewer with an immersive environment where he/she can focus on the object easily. Also, we present a preliminary design of the interface with motivation for application to a museum or an art gallery.

1. BACKGROUND AND PURPOSE

Many methods using eye tracking have been proposed for realizing interfaces for everyday device including digital signage and HMD. This is because eye tracking is able to reflect the user's interest, thought or intention effectively. Many of these visual interfaces rely on expensive eye tracking devices. Considering the cost and popularization, interfaces using such devices may be difficult to be proliferated. Gaze tracking was used to re-edit movies by Jain [1], and Zhang [2] proposed the visual interface using eye tracking to operate a software. On the other hand, Ono [3] introduced a means to estimate gaze direction from low resolution images. All of these have motivated us to develop audio interfaces based on eye-tracking.

Painting and music are deeply interrelated; it is indeed argued that for true understanding of an artwork, we need to listen to musical works of the time when it was actually drawn. Many attempts at synchronizing sound with pictures can be found (for example, [4]), though they focus primarily on the combination of an entire picture and a single audio data. In contrast, we herein intend to associate a suitable audio data with each local feature of a painting. Viewers can seamlessly enjoy the carefully chosen audio contents only by shifting their gaze from one feature to another. In this paper, we also present a system using such an interface, with a motivation for applying to an audio guide for museums.

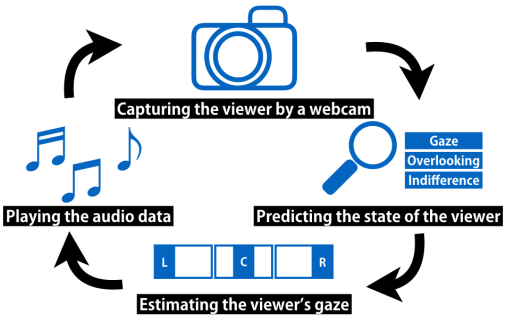

Figure 1: Processing loop adopted in our system

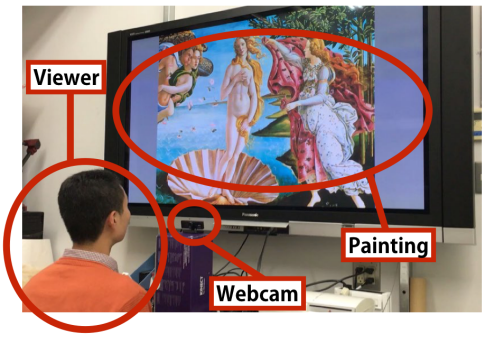

Figure 2: System in use

2. PROPOSED SYSTEM

Fig. 1 schematically shows the processing loop adopted in our system. A digital/physical painting and a webcam are placed in front of the viewer, as shown in **Fig. 2**. At the preprocessing stage, we have divided the painting into several areas and assigned a corresponding audio data to each of them. After this, the system captures the viewer and classifies the degree of his/her attention to the painting into three

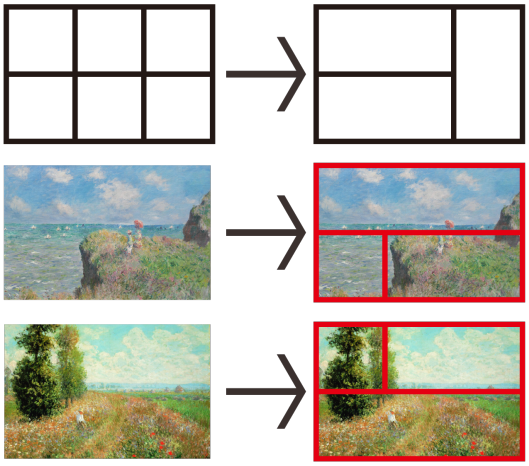

Figure 3: Examples of segmenting a target painting

types according as whether his/her face and/or eyes can be detected. Next, the system tracks the area the viewer is gazing at and plays the corresponding audio data. Every time the viewer's gaze changes, the system switches from one audio data to another in a cross-dissolve manner.

2.1. PAINTING SEGMENTATION

In our current prototype, it is assumed that any painting can be divided roughly into up to 6 (2 by 3) rectangular areas, as shown in **Fig. 3**. While the number of segmentation is limited, several patterns of segments derived from merging of the segments are expected to correspond roughly to representative feature configuration of paintings.

2.2. VIEWER'S STATE PREDICTION

In the prototype, we assume that the degree of the viewer's focus is different according to the distance between the user and the painting. It is classified into three states (**Fig. 4**): *indifference*, *overlooking*, and *gaze*. The state transition diagram is shown in **Fig. 5**.

Indifference means that the webcam cannot detect the viewer's face, implying that the viewer does not look at the painting.

Overlooking means that the webcam can only detect the viewer's face, implying that the viewer is just looking at the painting excursively.

Gaze means that the webcam can detect the viewer's face as well as his/her eyes, implying that the viewer is gazing at the painting.

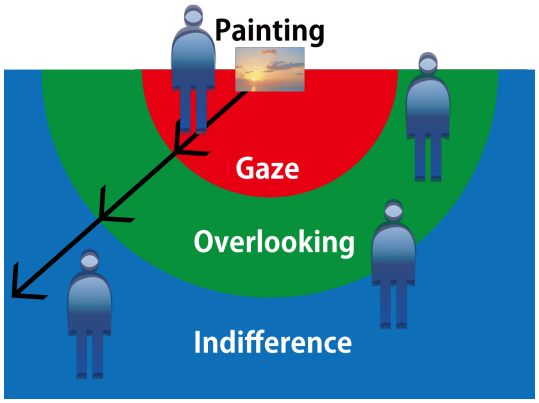

Figure 4: Distance-based classification of states

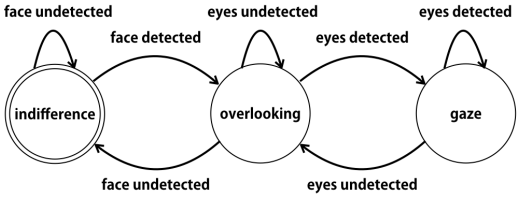

Figure 5: State transition diagram

2.3. GAZE ESTIMATION

Fig. 6 shows how the system finds the area the viewer is gazing at. The image from the webcam, which caputures the viewer in front of the painting, is converted into a corresponding gray scale image [5]. The system detects the viewer's eyes from the gray scale image, and trims one of his/her eyes and resizes it into the fixed size. The resulting image is then binarized using discriminant analysis method. For that end, the system seeks the pupil area—containing black pixels—at a rate of 10 times a second.

2.4. PLAYING AUDIO DATA

Fig. 7 shows how the system plays the audio data. At a preprocessing stage, we have divided the painting into several areas and assigned a corresponding audio data to each of them. In *Overlooking*, the system plays the global audio data representing the painting. In *Gaze*, the system plays the corresponding local audio data. Then, the system switches the audio data when the viewer's state has changed or when the area the viewer is gazing at has changed.

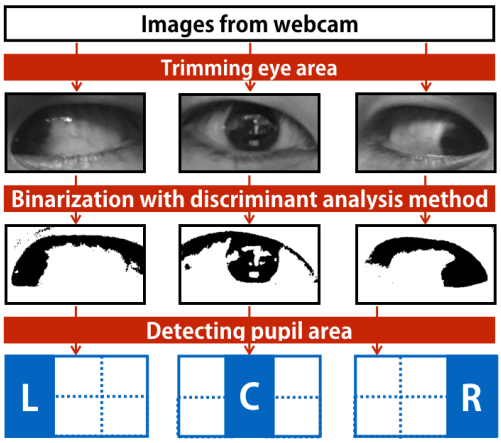

Figure 6: Flow of eye tracking

Figure 8: Experimental setting

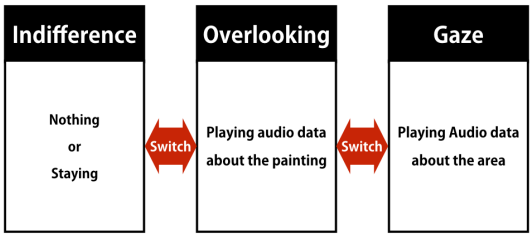

Figure 7: Scenario for playing audio data

3. RESULT AND EVALUATION

The present system has been implemented with OpenCV (http://opencv.org/) and OpenAL (http://www.openal.org/) and is running on a PC with an Intel Xeon E5540 2.53GHz CPU and a 12.0GB RAM with SONY SRS-X7 speaker.

We tried to apply our system to audio guide for an art museum. A digital painting and a webcam are placed in front of the viewer, as shown in **Fig. 8**. The thresholding distance from the painting to the viewer for the state of Gaze was set between 0.8 and 1.0m and the horizontal viewing angle of the webcam 30 degrees, as shown in **Fig. 9**.

The system works as expected. On trial, we used the system with five paticipants and compared with an existing manual audio guide. All of them reported that our system was more comfortable because they could enjoy the painting in a more immersive way.

4. ONGOING ISSUES

With the proposed system, viewers can seamlessly enjoy the carefully chosen audio contents only by shifting their gaze from one feature to another. Such a style of interface allowed the user to focus on paintnings more comfortably. In the current prototype, multiple users are not permitted. To relax the limit, we have to introduce face recognition to track each user's gaze. Also, it is currently assumed that any painting can be divided roughly into up to 6 rectangular areas. There are many actual paintings in accordance with the rule of thirds, so our system must be extended to support that by dividing the painting into at least 4 by 4 areas, as shown in **Fig. 10**.

In the proposed system, it is assumed that the painting has been divided into feature segments manually. It could be possible for the system itself to analyze the local features of the painting and to suggest adequate image segmentation candidates. Based on Arousal-Valence Space [6], the system is expected to associate the emotional aspects of each local feature with the audio properties.

Also, we can obtain fine grained painting–music correspondence by adjusting the audio data in terms of volume of music, speed of rhythm and pitch of tune according to the variance of local features of the painting. We are currently attempting to incorporate these functionalities into the system for the possible conference presentation.

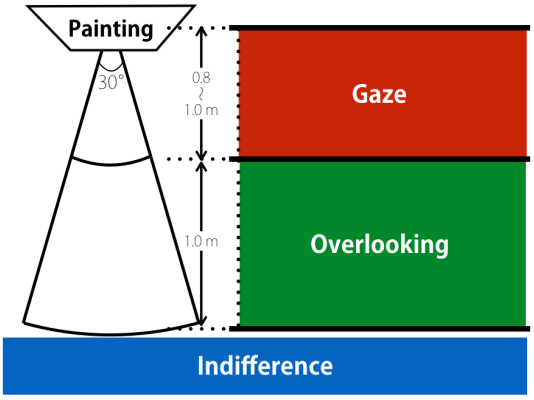

Figure 9: Distances for viewer's states

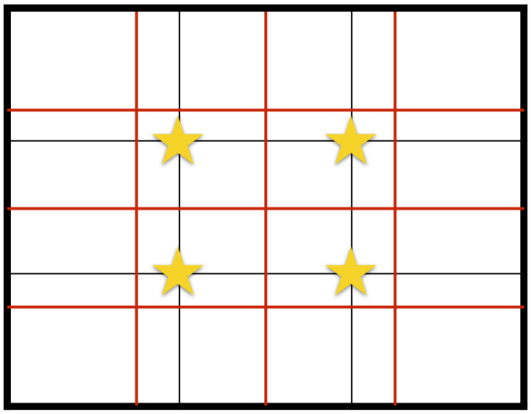

Figure 10: Rule of thirds

ACKNOWLEDGMENTS

This work has been in part supported by JSPS KAKENHI under the Grants-in-Aid for Challenging Exploratory Research No. 15K12034 and No. 16K12459.

REFERENCES

[1] Eakta Jain, Yaser Sheikh, Ariel Shamir, and Jessica Hodgins: "Gaze-driven video re-editing," *ACM Transactions on Graphics,* Vol. 34, No. 2, Article No. 21, 2015.

[2] Yanxia Zhang, Andreas Bulling, and Hans Gellersen: "SideWays: A gaze interface for spontaneous interaction with situated displays," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 851-860, 2013.

[3] Yasuhiro Ono, Takahiro Okabe, and Yoichi Sato: "Gaze estimation from low resolution images," *Advances in Image and Video Technology,* pp. 178-188, 2006.

[4] Liao Zicheng, Yu Yizhou, Gong Bingchen and Cheng Lechao: "Audiosynth: Music-driven video montage," *ACM Transactions on Graphics,* Vol. 34, No. 4, Article No. 68, 2015.

[5] Nobuyuki Otsu: "A threshold selection method from gray-level histograms," *Automatica,* Vol. 11, No. 285-296, pp. 23-27, 1975.

[6] James Russell: "A circumplex model of affect," *Journal of Personality Social Psychology,* Vol. 39, No. 6, pp. 1161-1178, 1980.