J-036

Video Content Analysis from Viewer's Perspective – Based on Pupil Size and Gazing Point

Kok-Meng Ong, Wataru Kameyama
GITS, Waseda University

Abstract— **the exponential increase in digital video volume has triggered the massive research effort in the area of video content analysis. Differ to conventional low-level audiovisual feature approach, this paper proposed to extract affective video content based on viewer's perspective. Experiments are carried out on 14 subjects and their pupil size and gazing point are recorded while watching a full-length motion picture. Affective features are extracted from pupil size and gazing point. The outcome shows that our proposal is able to represent the viewer's emotion and could be employed in video content analysis.**

I. INTRODUCTION

Digital video content has been increasing at exponential rate and the effort to appropriately retrieve desired content has become a tedious task. Researchers have proposed many content base video analysis methods [1]. Some of the works attempt to 'extract affect' from video content. In [2], an approach grounded upon psychology and cinematography is used to formulate the affective cue based on low level audio visual signals. [3] has proposed a framework that is based on the dimensional approach to affect. According to this approach, the video content is represented as a set of points in two dimensional emotion spaces, namely the arousal-valence curve. The video content is mapped into the arousal-valence curve by using low level features like shot duration, motion, and pitch. In [4] Hidden Markov Models (HMMs) is applied together with visual characteristic and camera motion at both the shot and scene level in an attempt to classify scenes depicting fear, happiness or sadness.

While these works have advanced research in affective classification, they are based on the video content itself. Due to the diversity in semantic interpretation from human to human, successful search of desired movie scene according to personal preference still remain an unresolved task. Therefore, we have proposed to analyze the video content based on the relationship between the video content and viewer's response. We monitor human response while they are watching movie by measuring their pupil size and gazing point.

II. THE SYSTEM ARCHITECTURE

We propose to relate the video content to human interest level. The system architecture is illustrated in Fig. 1. From the movie, low level features which constitute the audio and video component will be extracted. In this paper, the center of motion is extracted from the video component. To find out the human response, real time pupil size and gazing point of the experimental subject is recorded while they are watching the movie.

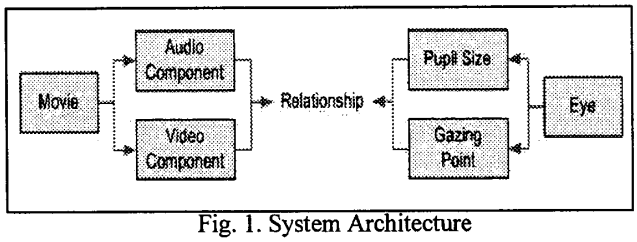

Fig. 1. System Architecture

III. CURRENT DEVELOPMENT

Our findings on the human response based on pupil size and gazing point has been published in [5]. In this paper, we report the finding on the relationship between video content and the previously published results.

A. ANALYSIS OF VIDEO CONTENT

Majority of motion estimation algorithm has been used to extract motion information at local basis. For example, the motion vector is determined for macroblocks in MPEG compression scheme.

However, human has remarkable ability to interpret complex scenes in real time and visual attention is focused on salient region, or 'focus of attention' [6]. Therefore, an algorithm to extract the center of motion is calculated in this paper as follow:

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

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

Where $P_{x,y|t}$ represent pixel at coordinate (x,y) at time t.

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

$$X = \sum_{x}^{Width} \sum_{y}^{Height} PD_{x,y} \times x \quad (2a)$$

$$Y = \sum_{x}^{Width} \sum_{y}^{Height} PD_{x,y} \times y \quad (2b)$$

3. Calculate the Euclidean Distance:

$$\text{Distance} = \sqrt{(X - X_{pupil})^2 + (Y - Y_{pupil})^2} \quad (3)$$

Where (X_{pupil}, Y_{pupil}) is the coordinate of gazing point.

The algorithm is depicted in Fig. 2 below. The steps are illustrated in the circles.

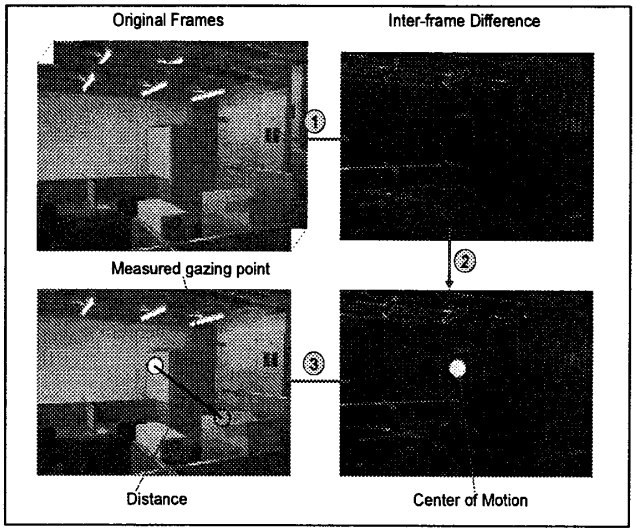

Fig. 2. Algorithm to Extract the Center of Motion

B. *REPRESENTATION OF HUMAN INTEREST*

In our previous work, several parameters have been extracted to model human interest [5]. In this paper, human interest, I(t) is modeled by pupil size.

$$I(t) = \sum_{t-T}^{t} P(t) \qquad (4)$$

I(t) represent the accumulated affect at time t, which is the summation of normalized pupil size, P(t) across time duration T. P(t) is given as follow:

$$P(t) = \frac{Pupil(t) - P_{baseline}}{P_{baseline}} \qquad (5)$$

Where Pupil(t) is the pupil size measurement at time t. $P_{baseline}$ is subject specific baseline pupil size value. It is obtained by calibration before the data collection.

IV. EXPERIMENT AND EVALUATION

14 volunteers participated in the experiment. The subjects were required to watch an Animation movie titled "Ratatouille" which running length is 102.25 minutes. The movie is cut into 15 scenes. Pupil size and gazing point were recorded from the participant's eye using VIS-EYE Measurement System [7] at 60 Hz sampling frequency.

Preliminary evaluation is carried out by finding the correlation between the inverse of Distance in equation (3) and the cumulative pupil size I(t) in equation (4). To perform the correlation, the data is truncated in to 30 second time windows. The results of the evaluation for every subject are shown in Fig. 3.

From the graph, we could see that there are certain areas that exhibit strong correlation for all the subjects (red circle) if compared to other regions of the movie. These areas could potentially be the interesting points of movie when the subject pupil size dilates while they are gazing at the center of the motion of the movie. While other areas (red square) exhibit subject specific characteristic, which show either strong positive or negative correlation individually at these locations.

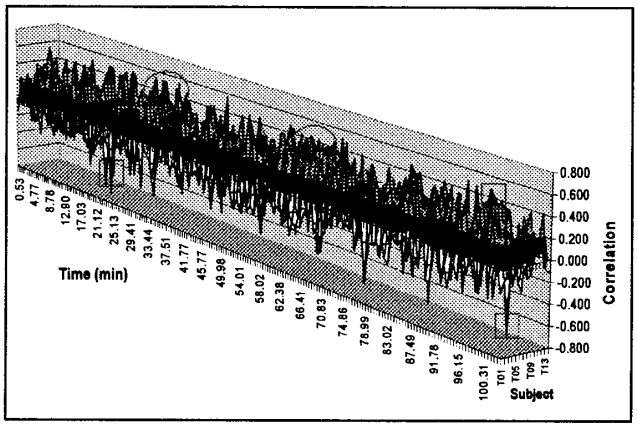

Fig. 3. Correlation between inverse of Distance and I(t) For each Subject

V. CONCLUSIONS AND FUTURE WORKS

This paper proposes to analyze the video content based on both the viewer's input and the video content. Based on the relationship between the motion center with pupil size, we could determine the interesting areas in the movie. The findings could be applied in wide application of video content analysis, and retrieval.

As our future works, we will continue evaluating other parameters that we could obtain from the pupil size and gazing point. In addition, more features of video content will be extracted to explore the relationship between the human emotions with low level features of audio visual signals.

ACKNOWLEDGMENT

This research has been conducted jointly with VIS Research Institute.

REFERENCES

[1] Ba Tu Truong and S.Venkatesh, "Video Abstraction: A Systematic Review and Classification", ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 3, No.1, Article 3, February 2007

[2] Hee Lin Wang and Loong-Fah Cheong, "Affective Understanding in Film", IEEE Transactions on Circuits and Systems For Video Technology, Vol.16, No.6, pp. 689-704, June 2006

[3] A. Hanjalic, Li-Qun Xu, "Affective Video Content Representation and Modelling", IEEE Transactions on Multimedia, Vol.7, No.1, February 2005

[4] H.B.Kang, "Affective content detection using HMMs." ACM Multimedia, pp.259-262, 2003.

[5] Kok-Meng Ong, Wataru Kameyama, "Quantifying Human Emotion during Video Watching based on Pupil Size and Gazing Point", No. 22-4, The Institute of Image Information and Television Engineers General Conference, Fukuoka, Japan, August 2008

[6] Laurent Itti, Christof Koch, and Ernst Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 20, No. 11, pp. 1254-129, November 1998

[7] http://www.visri/jp/english/index.html