視線の動きと映像のモーションセンターの動きの類似度に基づく映像要約手法の検討

オン コックメン　　亀山 渉
早稲田大学大学院 国際情報通信研究科
Email: ong_kok_meng@fuji.waseda.jp, wataru@waseda.jp

あらまし デジタルビデオコンテンツの急増につれ、視聴者の好みを基にしたビデオ視聴システムの需要が高まっている。そのため、本稿では視聴者の視線に基づいて、視聴者が興味ある映像部分を抽出する方法を提案する。具体的には、視聴者の視線の動きと映像のモーションセンターの動きを 1 秒間毎に空間ドメインと周波数ドメインで分析し、その動きの類似度を用いた映像要約手法を提案する。

キーワード 映像要約、視線、モーションセンター

On Consideration of Video Abstraction Based on the Similarity between the Gazing Point Movement and Movie Motion Center Movement

Kok-Meng Ong　　Wataru Kameyama
Graduate School of GITS, Waseda University
Email: ong_kok_meng@fuji.waseda.jp, wataru@waseda.jp

Abstract With the exponential increase in the volume of digital video content, it is crucial to efficiently consume the content based on viewer's preference. In this paper, we propose to identify the interesting parts of video based on viewer's gazing point. The viewer's gazing point movement is analyzed for every 1 second in both spatial and frequency domains. At the same time, the motion center of video is extracted and its movement is analyzed in spatial and frequency domains too. The video parts that the viewer is interested are obtained by finding the similarity between the gazing point movement and video motion center movement.

Keyword Video Abstraction, Gaze, Motion Center

1. Introduction

Digital video content has been increasing at exponential rate. This has fueled the research attention to the area of video abstraction in order to efficiently manage the huge amount of archive. Video abstraction is a mechanism for generating a short summary of video. A good video abstract will enable user to gain maximum information about the whole video in a specified time constraint [1].

In addition, the continuous dropping of digital storage media cost [2] and the advent of advance video compression technique [3] have brought terabytes of video into viewers' home. Furthermore, the available of multimedia delivery network has enabled Video-On-Demand consumers to seemingly enjoy countless content at their on pace. With the huge amount of digital video archive, a video abstract is crucial to aid users in managing the content. While abundance of work has been carried out in video abstraction algorithm based on the content itself [1], we believe that a combined approach of user feedback with the signal-analysis approach is more appropriate. This is because human response is subjective [4]. An abstract which is deemed important to a viewer might not be suitable for the other viewer.

In this paper, we propose to identify the interesting parts of video based on the similarity between the viewer's gazing point movement and the video motion center movement. The movement of viewer's gazing point and video motion center are analyzed for every second in both spatial and frequency domains. The interesting part of video is identified based on the level of similarity between the movement of gazing point and video motion center.

The rest of the paper is organized as follows. Our proposed approached is explained in detail in the following section. Section 3 presents our preliminary results and discussion. The paper is concluded in Section 4.

2. Approach

Our proposal to find the similarity between the gazing point movement and video motion center movement is illustrated in Fig. 1.

Human has remarkable ability to interpret complex scenes in real time and visual attention is focused on salient region, or 'focus of attention' [5]. Therefore, our approach is to match the movement of this visual attention to the video signals. We suggest that the when the viewer is attentive to the video content, their 'focus of attention' will match with the movement of the video.

From the video, the center of motion is extracted from the video component. To find out the human response, real time gazing point of the experimental subject is recorded.

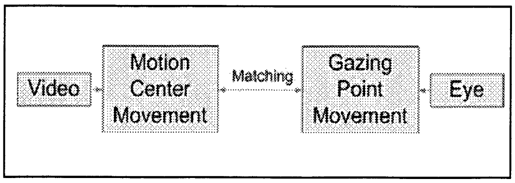

Fig. 1. Proposed Approach

2.1. Extraction of Video Motion Center Movement

The video's center of motion is extract based on the steps below [6]:

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

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

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

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

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

The extraction process is shown in Fig. 2.

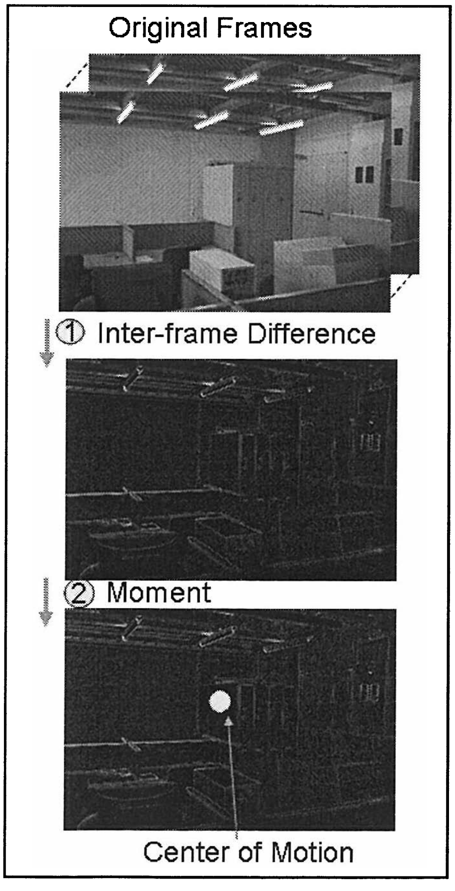

Fig. 2. Extraction of Video Motion Center

2.2. Extraction of Gazing Point Movement

The Gazing Point of viewer is obtained by the following method: A computer which host a standard 17 inches (42.5 cm) color monitor was used to administer the video viewing task and store responses. A second computer was used to control the eye tracking system and record the pupil size. Subject was seated in armless chair, facing the monitor with their head approximately maintaining a distance of 70 centimeter between the subject and the monitor. Gazing point was recorded from the participant's master eye using VIS-EYE Measurement System (Visual Interactive Sensitivity Research Institute Co. Ltd., Japan [7]). Gazing point was digitized at a 60Hz sampling rate and saved for offline processing. The pupil size was also recorded at the same time when the gazing point was measured. The measurement set up is shown in Fig. 3.

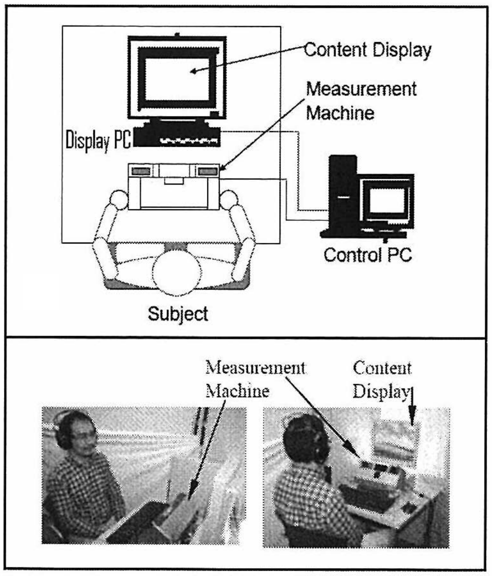

Fig. 3. Measurement Set-Up

2.3. Similarity

To obtain the movement for both gazing point and video motion center, images are drawn based on the past 1 second of data, which is made up of lines connecting 60 points in chronological order in the original dimension of the video, which is 320 × 240 pixels. The process is depicted in Fig. 4.

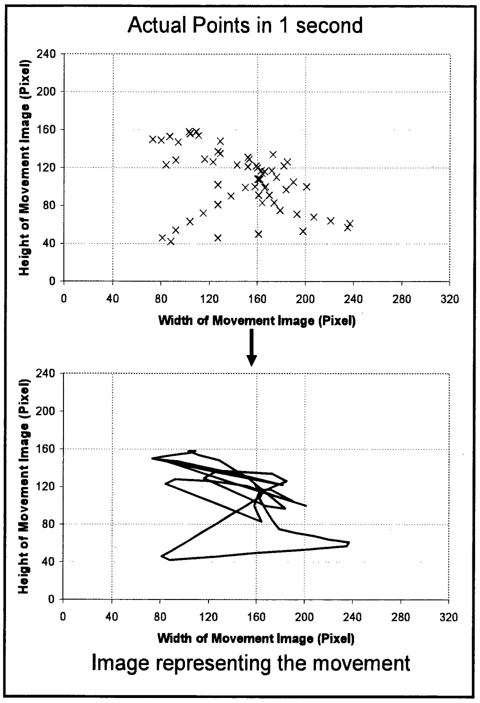

Fig. 4. Conversion of gazing points and video motion center into image

The image, which is taken as the representation of the movement is analyzed in both spatial and frequency domains. To convert to frequency domain, two-dimension Discrete Fourier Transform is applied:

$$f_{k_x k_y} = \sum_{n_x=0}^{N_x-1}\sum_{n_y=0}^{N_y-1} x_{n_x n_y} \exp\left(-\frac{2\pi i}{N_x}k_x n_x\right)\exp\left(-\frac{2\pi i}{N_y}k_y n_y\right) \tag{3}$$

Where $f_{k_x k_y}$ is the frequency component at coordinate (x,y), $x_{n_x n_y}$ is the pixel value component at coordinate (x,y), N_x and N_y are the total pixel at horizontal and vertical axis respectively.

To determine the similarity between the gazing point movement and video motion center movement, two matching methods are applied separately to both spatial and frequency domains. The matching methods are:

- Square Matching Difference:

$$R_{sq\_diff}(x,y) = \sum_{x,y}[T(x,y) - I(x,y)]^2 \quad (4)$$

- Correlation Matching:

$$R_{ccorr}(x,y) = \sum_{x,y}[T(x,y) \bullet I(x,y)]^2 \quad (5)$$

Where $T(x,y)$ and $I(x,y)$ are the image obtained from the gazing point and motion center respectively.

The overall process is depicted in Figure 5.

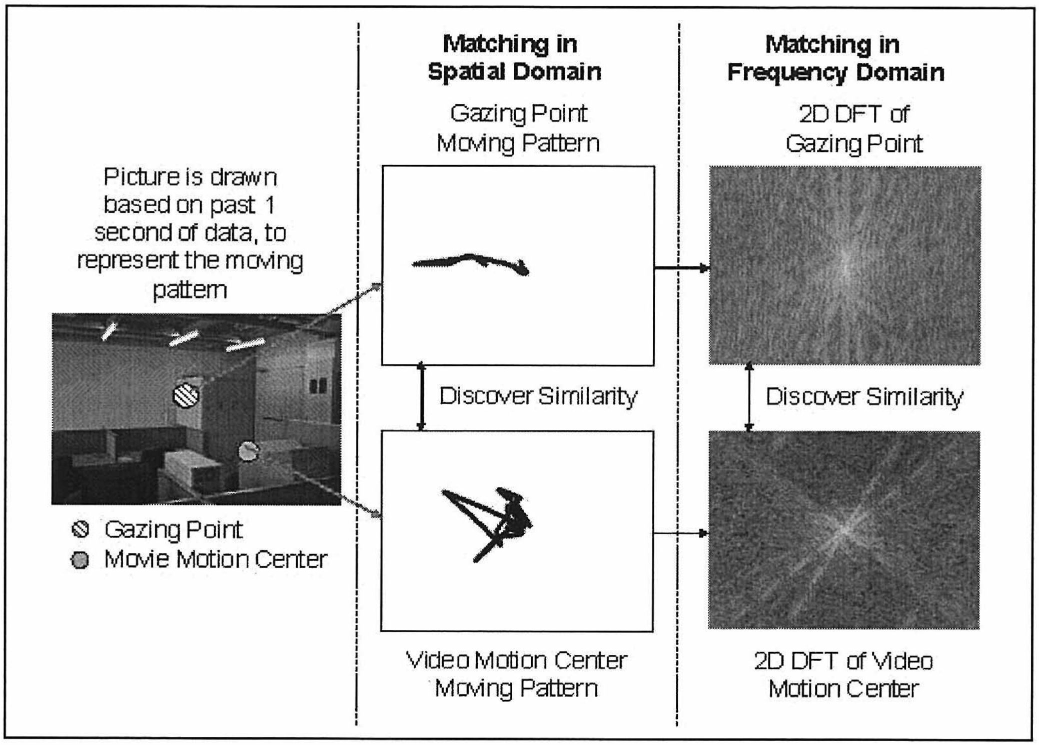
Matching in Spatial Domain
Gazing Point Moving Pattern
Matching in Frequency Domain
2D DFT of Gazing Point
Picture is drawn based on past 1 second of data, to represent the moving pattern
Discover Similarity
Discover Similarity
Gazing Point
Movie Motion Center
Video Motion Center Moving Pattern
2D DFT of Video Motion Center

Fig. 5. The Matching Method

3. Results and Discussion

Six subjects were requested to participate in the experiment by watching a video clip for approximately 6.5 minutes while their gazing point movement is recorded. The video is extracted from Animation Movie, Ratatouille. The proposed analysis method was carried out and the similarities with the video motion center movement are analyzed in spatial and frequency domain.

In the spatial domain, the Correlation Matching method in equation (5) is applied to images extracted from the gazing point and video motion center to find the similarity. The average result for all the subjects is depicted in Fig. 6.

For analysis in the frequency domain, the images are first converted by equation (3). Then, the Square Matching Difference is calculated based on equation (4). The average result for all the subjects is depicted in Fig. 7.

From the figures, it is observed that there are certain areas in the video clip that both the movements match and vice versa. To achieve video abstraction, a threshold could be applied to trim the video to desired length of summary. The threshold could be adjusted according to the desired application. In addition, key frames could be generated by extracted frames that are corresponding to the peaks in the graph.

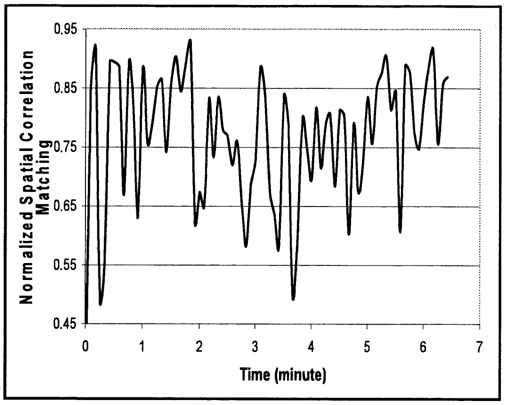

Fig. 6. Correlation Matching in Spatial Domain

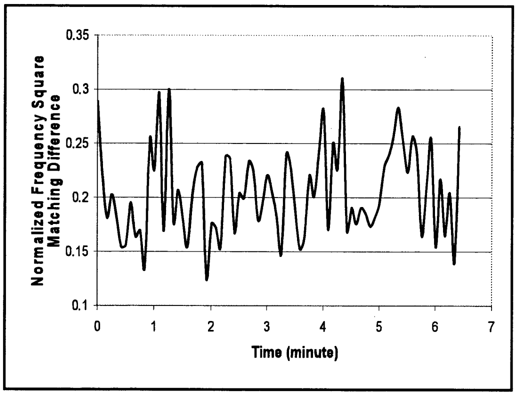

Fig. 7. Square Matching Difference in Frequency Domain

In addition, the similarity between the two movements for each individual subject are obtained and depicted in Fig. 8. For the purpose of better visualization, the graph shows the individual correlation matching in spatial domain only for 3 different subjects.

Two different views can be observed from the results. First of all, at certain location along the video, most of the subjects are having good matching for both the movements. These areas are reflected on Fig. 6., where the average value is high. These video areas could possibly be the area where most of the subjects are interested in, and could be used to generate a general abstraction for other users which has no gazing point data recorded.

On the other hand, it is clearly observed in Fig. 8. that at some point of time, the similarity level between the two movements for every subject is different. These are the areas where a generalized approach could not be applied and highly personalized approached is preferred if the gazing points data are available.

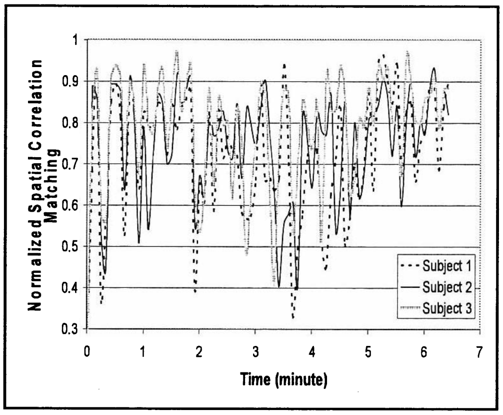

Fig. 8. Correlation Matching in Spatial Domain for 3 Subjects

4. Conclusion and Future Work

In this paper, a video abstraction method based on the similarity between viewer gazing point movement and video motion center movement is proposed. The similarity is analyzed in both spatial and frequency domains. From the experiment results, it is observed that the similarity level of these movements varies during video watching. Based on this observation, we propose the generating of video abstract by extracting the parts of video when the similarity is high. In addition, we have observed that individual responses differently even to the similar video content.

Future works need to be carried out to improve the extraction of video motion center. The current motion center extraction is based on moment calculation which is lacking in identifying conversation scene between the characters in video. We are currently working on applying advance image processing technique to extract human face and to generate the movement characteristic based on the video.

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 Transaction on Multimedia Computing, Communications and Applications, Vol. 3, No.1, Article 3, February 2007.
2. Schaller, R.R., “Moore’s Law: Past, Present, and Future”, IEEE Spectrum, Vol. 34, Issue 6, pp. 52-59, Jun 1997
3. Mohammed Ghanbari, “Standard Codecs: Image Compression to Advance Video Coding”, The Institute of Electrical Engineers, 2003
4. R.W. Picard, “Affective Computing”, The MIT Press, 2000
5. Laurent Itti, Christof Koch, and Ernst Neibur, “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 – 1259, November 1998
6. 土井滋貴, “はじめての動画処理プログラミング”, CQ 出版社, 2007
7. Visual Interactive Sensitivity Research Instituted Co. Ltd. http://www.visri.jp/english/index.htm