Signal Processing: *Image Communication* 39 (2015) 405–417

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Temporal resolution vs. visual saliency in videos: Analysis of gaze patterns and evaluation of saliency models

Manri Cheon [a,b], Jong-Seok Lee [a,b,*]

[a] *School of Integrated Technology, Yonsei University, 406-840 Yeonsu-gu, Incheon, Republic of Korea*
[b] *Yonsei Institute of Convergence Technology, Yonsei University, 406-840 Yeonsu-gu, Incheon, Republic of Korea*

ARTICLE INFO

Available online 19 June 2015

Keywords:
Temporal scalability
Eye-tracking
Frame rate
Perception
Quality of experience
Saliency model

ABSTRACT

Temporal scalability of videos refers to the possibility of changing frame rate adaptively for efficient video transmission. Changing the frame rate may alter the spatial location that the viewers pay attention in the scene, which in turn significantly influences human's quality perception. Therefore, in order to effectively exploit the temporal scalability in applications, it is necessary to understand the relationship between frame rate variation and visual saliency. In this study, we answer the following three research questions: (1) Does the frame rate influence the overall gaze patterns (in an average sense over subjects)? (2) Does the frame rate influence the inter-subject variability of the gaze patterns? (3) Do the state-of-the-art saliency models predict human gaze patterns reliably for different frame rates? To answer the first two questions, we conduct an eye-tracking experiment. Under a free viewing scenario, we collect and analyze gaze-paths of human subjects watching high-definition (HD) videos having a normal or low frame rate. Our results show that both the average gaze-path and subject-wise variability of the gaze-path are influenced by frame rate variation. Then, we apply representative state-of-the-art saliency models to the videos and evaluate their performance by using the gaze pattern data collected from the eye-tracking experiment in order to answer the third question. It is shown that there exists a trade-off relation between accuracy in predicting the gaze pattern and robustness to frame rate variation, which raises necessity of further research in saliency modeling to simultaneously achieve both accuracy and robustness.

1. Introduction

Nowadays, video content delivery services over networks have become popular and a significant amount of contents are consumed through online video services. Many TV broadcasters have switched their broadcasting systems from analog to digital and now provide TV services using Internet, e.g., IPTV. In addition, many online video services such as YouTube, Netflix and Vimeo provide high quality videos over networks. As the technology is advanced and popularized, high quality videos are made, distributed, and viewed easily by public users as well as experts. In particular, high definition (HD) video contents are popularly consumed in many video applications, which achieve improved visual quality at the expense of an increased volume of data. As this trend continues, it can be predicted that the video traffic will continuously increase. Therefore, there exist challenges in these video services, including the limit in the network capacity and user heterogeneity in terms of the network environment and terminal capability.

[*] Corresponding author at: School of Integrated Technology, Yonsei University, 406-840 Yeonsu-gu, Incheon, Republic of Korea.
E-mail addresses: manri.cheon@yonsei.ac.kr (M. Cheon), jong-seok.lee@yonsei.ac.kr (J.-S. Lee).

http://dx.doi.org/10.1016/j.image.2015.05.010
0923-5965/© 2015 Elsevier B.V. All rights reserved.

Video scalability provides flexible solutions to such challenges, i.e., the data rate and quality parameters of a scalable video can be adapted without the necessity of re-encoding of the video data [1]. It is supported by recent video compression standards, e.g., the scalable extension of H.264/AVC (SVC) [2] and the scalable extension of HEVC (SHVC) [3]. In addition, the recently standardized dynamic adaptive streaming over HTTP (DASH) technique [4] provides an efficient framework to implement video streaming systems based on video scalability [5]. There are several dimensions of video scalability, e.g., temporal, spatial, and quality scalability. Among these, the temporal scalability refers to the possibility of changing frame rate adaptively. It has been used singly or in combination with other scalability dimensions for video transmission in recent studies [6,7].

In general, changing the frame rate affects human's visual perception significantly. Thus, it is important to understand the effect of frame rate variations on viewer's perception in order to exploit the temporal scalability effectively. The effects of the video frame rate on human performance were investigated by Chen and Thropp [8]. In the study, researches about the effects of frame rates on perceptual performance were reviewed, from which it was concluded that a threshold of the low frame rate for conducting psychomotor and perceptual tasks is around 15 Hz.

In particular, we focus on the influence of the frame rate change on the visual attention in this study. Changing the frame rate alters perceived motion information, which consequently affects visual attention. In the study of Itti et al. [9], it was shown that the bottom-up visual attention mechanism is influenced by motion information in the given visual stimulus.

The visual attention has been considered as important in many applications related to perceptual quality, video compression, computer vision, etc. Quality perception of viewers is highly dependent on where they pay attention in the scene. It was shown that visual artifacts are likely more annoying in a salient region attended by viewers than those in other areas by Ninassi et al. [10]. You et al. [11] proposed a quality metric based on balancing the influence of the entire visual stimuli and attended stimuli on the perceived video quality. Video compression can also exploit different visual sensitivity of humans for attended and unattended regions, which is referred to as perceptual video compression [12]. Machine vision is also often based on the visual attention mechanism in order to simulate human's capability of visual scene understanding, visual search, etc.; for instance, a model integrating top-down and bottom-up attention for fast object detection was proposed by Navalpakkam and Itti [13]. It is apparent that, if the video frame rate variation significantly affects the visual attention, such effects need to be considered in designing the aforementioned applications based on the visual attention.

Gulliver and Ghinea [14] performed an eye-tracking study related to the frame rate and perceptual quality. They noted that the gaze-path is not significantly affected by frame rate variations, which was based on the observation that the median gaze-path over viewers is consistent independently of the presentation frame rate varied among 5 fps, 15 fps, and 25 fps. However, it is questionable whether such conclusion is still valid for the up-to-date video consumption environment (such as HD or ultra high-definition (UHD)) that is largely different from that considered in their study. With the advances in the video technology, the screen size, display resolution, and video content resolution have ever increased, and accordingly, the horizontal viewing angle has also increased [15]. Consequently, human's perceptual patterns of video content has changed; it was shown that the perception of presence (immersion and perceptual realism) is affected by the screen size [16]. In particular, the influence of the visual attention on perception of video content became more and more prominent. Furthermore, while only simple examination of the median gaze-path was performed in the study of Gulliver and Ghinea [14], deeper analysis in various aspects is needed to understand the perceptual mechanism better. In fact, our preliminary study demonstrated that the frame rate variation changes the gaze-path in an average sense over subjects [17].

For the applications using the gaze information, it is important to imitate the human's gaze pattern precisely by using a visual saliency model [18]. In the literature, many saliency models have been developed [19]. However, their applicability across different video frame rates has not been studied before. Ultimately, it is desirable for a saliency model to be able to accurately predict the human gaze pattern across different video frame rates consistently, which imposes two (possibly conflicting) criteria: accuracy and robustness (or consistency). Then, the model can be used without modification across applications involving different frame rates or will work well in an application having the possibility of frame rate variations during its operation.

In this paper, we will answer the following three research questions: (1) Does the frame rate influence the overall gaze patterns (in an average sense over subjects)? (2) Does the frame rate influence the inter-subject variability of the gaze pattern? (3) Do the state-of-the-art saliency models predict human gaze patterns reliably for different frame rates? To answer the first two questions, we conduct an eye-tracking experiment. Under a free viewing scenario, we collect and analyze gaze-paths of human subjects by watching HD videos having a normal or low frame rate. Then, we apply representative state-of-the-art saliency models to the videos and evaluate their performance by using the gaze pattern data collected from the eye-tracking experiment, which will answer the third question.

The rest of the paper is organized as follows. The following section explains how the eye-tracking experiment was designed and conducted, and presents the results of the experiment. Section 3 is devoted to the evaluation of saliency models based on the collected gaze pattern data. Finally, conclusions are given in Section 4.

2. Eye-tracking experiment

2.1. Test sequences

Eighteen HD sequences that have a frame size of 1920×1080 pixels were selected and used for the eye-tracking experiment. These sequences originally have a frame rate of 25 fps or 30 fps, which we refer to normal frame rate (NFR). Most of the sequences were obtained