P−2.15

ビデオストリーミングのためのフレーム間特徴マップを用いた視点予測
Estimating Visual Attention using Inter-Frame Saliency Map Analysis for Gaze-based Video Streaming

馮 雲龍[†] チョン ジーン[‡] 計 宇生[‡]
Yunlong Feng[†] Gene Cheung[‡] Yusheng Ji[‡]

総合研究大学院大学[†] 国立情報学研究所[‡]
The Graduate University for Advanced Studies[†] National Institute of Informatics[‡]

Abstract: A viewer's ability to perceive details deteriorates as function of the viewing angle away from his eye gaze focal point. Thus, a smart video coding scheme can allocate more bits to the spatial area enclosing his gaze focal point (region of interest ROI) and fewer bits elsewhere, without any degradation in perceived visual quality. In a server-client streaming scenario, however, a viewer's eye gaze must be predicted one round-trip time (RTT) into the future to avoid delay between ROI-based bit allocation at server and viewer’s gaze movements at client. In our previous work, we devised a Hidden Markov Model (HMM) to predict a viewer's gaze focal point one RTT into the future, using real-time collected eye-gaze data as input. However, HMM parameters must be trained a priori using acquired gaze data. In this work, leveraging on recent research in visual saliency maps, we derive HMM statistics by analyzing saliency maps offline. Our analysis can also detect an abrupt change in gaze statistics, so that a video can be appropriately segmented into clips, each with stationary gaze statistics.

1 Introduction

It is known that a viewer's ability to perceive visual details deteriorates as a function of the viewing angle away from his eye gaze focal point. Thus, a smart video coding scheme can allocate more bits to the spatial area enclosing his gaze focal point (region of interest ROI) and fewer bits elsewhere, without any degradation in perceived visual quality. In a server-client streaming scenario, however, given a viewer's eye gaze can change often, even if gaze is tracked in real-time at client, the response in reallocation of bits at server will necessarily suffer a round-trip time (RTT) delay. In our previous work [1], we devised a Hidden Markov Model (HMM) to predict a viewer's gaze focal point one RTT into the future, using real-time collected eye-gaze data as input, for smart bit allocation at server. Our experiments show that bit rate can be reduced by up to 21% without noticeable visual quality degradation when end-to-end network delay is as high as 200ms. HMM must be trained off-line to derive suitable parameters for each video clip, however, leading to a complex process.

In this work, leveraging on recent research in visual saliency maps [2]—estimation of viewers' ROI via synthesis of detected low-level features in the video like motion and flickers—we derive HMM statistics by analyzing saliency maps offline without collecting eye-gaze data. Our methodology is simple and intuitive, and has been shown to be effective for a range of videos with drastically different gaze statistics. Our analysis can also detect an abrupt change in gaze statistics (by calculating the Kullback-Leibler divergence of neighboring frame statistics), so that a video can be appropriately segmented into clips, each with stationary gaze statistics.

2 HMM for Eye-gaze Prediction

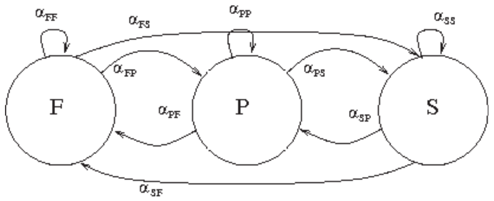

Fig.1: HMM for eye-gaze movement

We first discuss how we model eye gaze of a video viewer using a *hidden Markov model* (HMM). An HMM models transitions of sequential states X_n’s in discrete time, where X_n is the state variable at time n. Each X_n can take on one of three possible latent states. State **F** (*fixation*) models the case when eye gaze is fixated at a stationary object. State **P** (*pursuit*) models the case where gaze follows motion of a moving object. State **S** (saccade) models the case where gaze rapidly transitions from one fixation point to another. Broadly speaking these are the three major types of eye movements. See Fig.1 for an illustration.

An HMM is Markovian in that the determination of state variable X_{n+1} at time $n+1$ depends solely on the value of X_n of previous time n. In particular, given $X_n=i$, the probability of $X_{n+1}=j$ is represented by *state transition probability* $\alpha_{i,j}$ of switching from state i to j. $\alpha_{i,j}$'s are the HMM parameters we sought to derive via visual saliency map analysis next.

3 Analysis of Saliency Maps

We first compute visual saliency maps for all video frames using methodology in [2]. We then normalize each one, so the sum of all saliency values in a frame equals to one. We then find a set of *saliency objects* in each map: spatially connected regions with per-pixel saliency value larger than a pre-defined threshold τ_s. As an approximation, we assume these are the only objects a viewer will observe in the given frame. A viewer may of course have gaze location outside of these saliency objects; we assume such occurrence means the viewer is in the process of switching from one saliency object to another; i.e., he is in saccade state **S** at this frame time.

We can establish correspondence among saliency objects in consecutive frames by matching RGB pixel values of candidate objects. We can then use motion information of corresponding saliency objects in consecutive frames to label each object either as a stationary or moving object. A viewer's gaze following a saliency object that is stationary or moving will be in state **F** or **P**, respectively.

Having identified saliency objects across frames, we now derive state transition probabilities $\alpha_{i,j}$'s in the eye gaze HMM. Essentially, we write equations to establish consistency in probabilities during HMM state transitions from objects in frame t to objects in frame $t+1$, relative to the sizes of saliency objects in two frames. See [3] for details.

4 Segmenting Video via KL Divergence

We can determine how gaze statistics are changing in a video by computing the Kullback-Leibler (KL) divergence, treating consecutive motion-compensated saliency maps as probability density functions. If the computed KL divergence exceeds a certain threshold τ_{KL}, then we can divide the video into segments of different gaze statistics.

5 Results

We compare the computed steady probabilities π_S for state **S** (a measure of how often a viewer switches gaze location) using saliency map analysis with ones trained using gaze traces for two MPEG test sequences, *Silent* and *Table*. In Table 1, we see that the two sets of probabilities are similar, validating our saliency map analysis approach.

Table 1: HMM Parameter Comparison.

	π_S
Gaze trace (silent)	0.0628
Saliency map (silent)	0.0788
Gaze trace (table)	0.4318
Saliency map (table)	0.4767

We have also compute the KL divergence for a 300-frame video that is a concatenation of 3 100-frame segments: 100 frames of *Silent*, 100 frames of *Table*, and 100 frames of *Silent*. Because of the change in content at frame 101, and 201, we expect computed KL divergence using computed motion-compensated saliency maps to have large values at these locations. In Fig. 2, we indeed see that this is the case, showing the utility of using KL divergence to divide a video into different segments of different statistics.

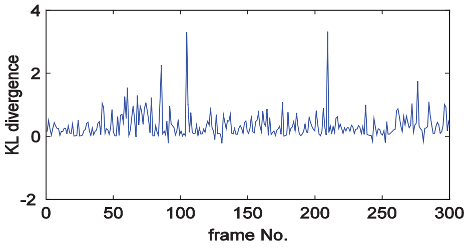

Fig.2: Kullback-Leibler(KL) divergence

References

[1] Y. Feng *et al.*, "Hidden Markov Model for Eye Gaze Prediction in Networked Video Streaming," *IEEE ICME*, Barcelona, Spain, July 2011

[2] L. Itti *et al.*, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on PAMI*, vol.20, no.11, Nov. 1998.

[3] Yunlong Feng *et al.*, "Video Attention Deviation Estimation using Inter-Frame Visual Saliency Map Analysis," accepted to *IS&T/SPIE VIPC*, Burlingame, CA, January 2012.

国立情報学研究所
〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: {fengyl,cheung, kei}@nii.ac.jp