Can Visual Scanpath Reveal Personal Image Memorability? Investigation of HMM Tools for Gaze Patterns Analysis

Waqas Ellahi, Toinon Vigier, Patrick Le Callet

2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)

Can Visual Scanpath Reveal Personal Image Memorability? Investigation of HMM Tools for Gaze Patterns Analysis

Waqas Ellahi, Toinon Vigier, Patrick Le Callet
LS2N UMR CNRS 6003, Université de Nantes, Nantes, France
firstname.name@univ-nantes.fr

.

***Abstract*—Visual attention has been shown as a good proxy for QoE, revealing specific visual patterns considering content, system and contextual aspects of a multimedia applications. In this paper, we propose a novel approach based on hidden markov models to analyze visual scanpaths in an image memorability task. This new method ensures the consideration of both temporal and idiosyncrasic aspects of visual behavior. The study shows promising results for the use of indirect measures for the personalization of QoE assessment and prediction.**

***Index Terms*—scanpath, eye movements, visual attention, hidden markov model, memorability**

I. Introduction

Visual attention (VA) is the process that concentrates the processing resources on an appropriate visual information. VA computational modeling efforts has been mostly focuses on visual saliency, providing the probability for each visual location to attract attention. Such models have been successfully used to improve Quality of Experience (QoE) in many multimedia applications [1], [2]. However, saliency models do not reflect the dynamic (saccadic) nature of visual scanpaths [3], Consequently, more efforts have been made to analyse and model visual scanpaths, notably, using probabilistic approaches considering that eye movement parameters are random variables generated by underlying stochastic processes [4]–[7]. It has been demonstrated that visual scanpath patterns can exhibit idiosyncratic behavior when watching visual stimuli, especially for high level cognitive, task [8], [9]. As QoE evaluation for multimedia application implies to assess the impact of contextual and individual factors [10], visual scanpath patterns look relevant candidate to direct measure of individual aspect of user experience.

In this paper, we consider one particular high level cognitive factor of visual experience: image Memorability (IM). IM understanding can be useful to improve user experience in many use cases, such as content sorting, selection or recommendation [11]. Computational modeling of IM is a topic that has recently raised and different IM prediction models have been developed using computer vision and machine learning techniques [12]–[14]. In these efforts, IM is defined as the probability of an image to be recognized a few minutes after a single view. Ground truth have been collected for short-term memorability (estimation after few minutes of the memorization process) and long-term memorability (24-72 hours after the memorization process) [11], [15], [16]. While IM models usually predict mean memorability opinion score, it has been shown that emotional and individual biases are important in memorability tasks [15] making IM dimension a relevant candidate to study individual aspect of user experience.

978-1-7281-5965-2/20/$31.00 ©2020 IEEE

Consequently, this paper explores the potential of Hidden Markov Models (HMM) based framework to study the link between behavior of eye movement and image memorability for individual user. To our knowledge, such relationship between IM and visual scanpath has never been studied before. The rest of this paper presents HMMs based technique for visual scanpath analysis, dataset and methodology and results of the analyses.

II. HMM tools for visual scanpath analysis

HMM is a statistical model that can be used to observe the outcome with help of set of stochastic process that produces the sequence of observation. The outcome is predicted by estimating the transition among different states. The advantages of HMM are data-driven and handling time series information. Additionally, it can integrate influences stemming from both bottom-up and top-down mechanisms, as well as viewing biases into a single mode. In the proposed study, estimated regions of interest (ROIs) of the image are set as the hidden states of HMM in order to observe the eye position. Following the technique used in Chuk et al. [5], the transition matrix in our case represents the saccadic movements of the observer. The emission densities (the distribution of fixations in each ROI) are modeled as two-dimensional Gaussian distributions as shown in Fig. 2. The initial state of the model, i.e., the probability distribution of the first fixation, is modeled with the prior values.

The variational approach to Bayesian inference enables simultaneous estimation of model complexity and model parameters [17], which automate the process of finding the number of states (K) in HMM model [5]. An HMM is trained on eye movement data which enables to compare and visualize the gaze behavior of different groups of observers, in different experimental conditions. The individual HMMs of the similar gaze pattern behaviour are clustered into joint HMM with

a variational hierarchical expectation maximization (VHEM) algorithm proposed by Coviello et al. [18], [19]. For each cluster, the VHEM algorithm produces a representative HMM, which summarizes the common ROIs and transitions in the cluster.

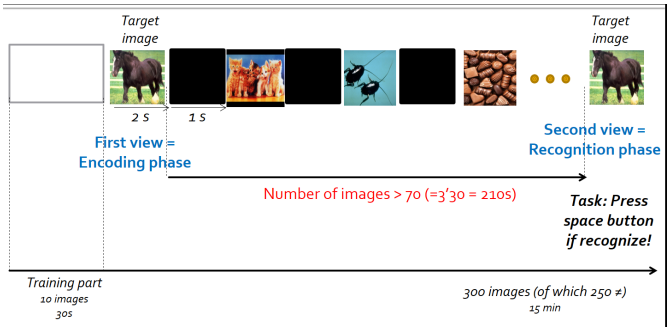


Fig. 1: Experimental protocol for memorability assessment.

III. Application on individual image memorability prediction

A. Dataset and experimental setup

We used the IM dataset with eye data described in [15]. This dataset has been collected using forty-nine subjects. The test employed 625 images which were randomly selected from the International Affective Picture System (IAPS) [20] presented on a 40-inch monitor. This experiment was split in three sessions: short-term memorability assessment, long-term memorability assessment and emotional assessment. During each session, gaze data was recorded using SMI RED eye tracker. In this paper, only the data of the first session (short-term memorability) are used. During this session, the observer performed a memory encoding task (encoding phase) in which they were presented a stream of 50 target images twice and recognized the target images out of a sequence of 200 (recognition phase) images (see Fig. 1). Each target image was seen twice by 14 to 17 observers. In the following paper, eye data recorded during the first viewing and second viewing is labeled as encoding phase and recognition phase respectively. In addition, scanpaths are clustered and labeled in two classes, correct and wrong, i.e. if observers recognized the target images during the second viewing or not.

B. Methodology for individual IM recognition based on gaze data

Initially, HMM observation for each participant is estimated to record the pattern in eye movements. The HMM responses are categorized into two labels corresponding to correct and wrong recognition of target images during the second phase by the participants. Indeed, we observe, for some content, different eye patterns for correct and wrong clusters in the different phases Fig. 2. The eye patterns for correct and wrong responses in encoding phase seem similar to the patterns in the recognition phase, whereas they seem different for each other. Considering this observation, individual responses placed in each category (correct/wrong) and each phase (encoding/recognition) are then clustered into single Joint HMM by using VHEM. Several scenarios are then applied as shown in Fig. 3 and detailed below.

1) Analysis I: HMM clustering in encoding and recognition phase: In the first analysis (Scenarios (a) and (b) in Fig. 3), we are testing if IM can be predicted, in encoding and recognition phases, by classifying the scanpaths of each phase in two clusters, correct and wrong. For each phase, the classification is performed by estimating the log-likelihoods of scanpaths with correct HMMs and wrong HMMs by following the leave-one-out validation approach, i.e., one observer scanpath is used as a test data and HMM is trained on remaining data. The accuracy of the proposed scheme is measured by counting the number of times a class is correctly guessed divided by the total number of iteration (N).

2) Analysis II: Comparison of scanpath in recognition task with HMM models of encoding phase: In this second analysis (Scenarios (c) and (d) in Fig. 3), the main idea is to compare correct and wrong scanpaths during the recognition phase with correct and wrong scanpaths during the encoding phase. Towards that, joint correct and wrong HMMs are computed for the encoding phase (first viewing of the target images). Later on, the scanpaths recorded for target image recognition phase are used as test data and likelihoods are estimated for both categories. The analysis yields two vectors of mean log-likelihoods of the correct and wrong HMMs respectively. we then examined whether the two vectors are significantly different using t-test. This procedure is repeated on the wrong fixation sequences of the recognition data.

IV. Results and Discussion

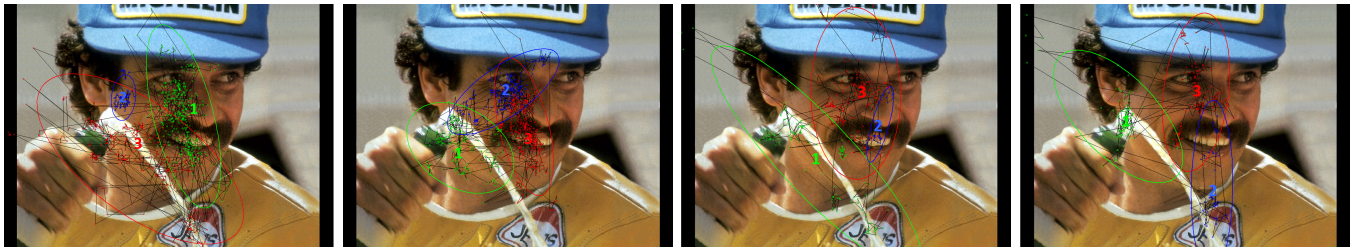
In our experiments, the maximum value of parameter K is set to 3. In addition, only images which have at least five responses for both correct and wrong recognition were selected, which leads to a total of 61 different content tested on 14 to 17 different observers.

A. Analysis I

Fig. 4 shows the accuracy of the proposed system for encoding and recognition phases. The proposed scheme achieves 44.32% and 58.03% accuracy for the encoding and recognition data respectively with leave-one-out validation. It is observed that the eye data pattern for recognition phase is more differentiate as compared to the encoding phase.

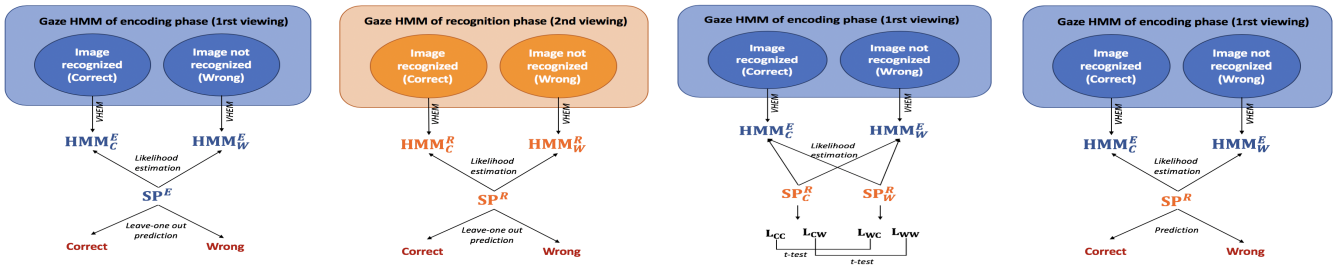
B. Analysis II

From this second analysis (see Section III-B2), we obtain, for correct (resp. wrong) joint HMM using VHEM of encoding phase, two vectors of log-likelihoods values, L_{CC} (respectively L_{WC}) and L_{CW} (resp. L_{WW}) for correct and wrong scanpaths of recognition phase. Pairwise t-test is then used to compare L_{CC} and L_{CW}, as well as L_{WC} and L_{CW}. Results show 20 significantly different pairs for L_{CC} and L_{CW}, i.e. comparison with correct encoding scanpaths, and 5 significantly different pairs for L_{WC} and L_{WW}, i.e.



(a) Correct responses for encoding phase (b) Correct responses for recognition phase (c) Wrong responses for encoding phase (d) Wrong responses for recognition phase

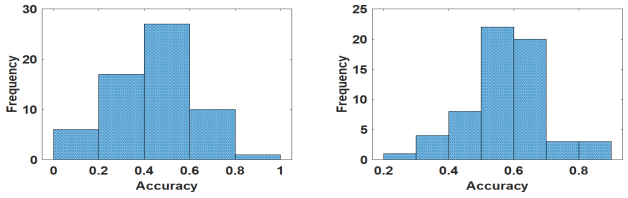
Fig. 2: Gaze data plot and HMM model output of different response for different phases



(a) Encoding phase classifier (b) Recognition phase classifier (c) Statistical comparison (d) Cross phase classifier

Fig. 3: Classification and analysis methodology (SP denotes scanpath; L denotes log-likelihood scores)

comparison with wrong encoding scanpaths. Four of the five different pairs for the second scenario are also significantly different for the first one. These results show a tendency to regularity between correct gaze pattern in encoding and recognition phases.



(a) Encoding Phase scanpath accuracy (b) Recognition Phase scanpath accuracy

Fig. 4: Histogram of accuracy score.

From these encouraging results, we test a new memorability prediction scenario based on the estimation of log-likelihoods of scanpaths in recognition phase with correct and wrong HMMs in encoding phase (see fig 3(d)). In this scenario, training set is directly based on encoding data and testing set is the recognition phase scanpaths. This scheme obtains an accuracy if 61.48% in this analysis, which is slightly better than the two other scenarios.

C. Limitations and future works

Results, from statistical analysis and classification, tend to show an impact of image content for the study and the prediction of image memorability from gaze data. Future work would be to better characterize the content for which differences between correct and wrong eye patterns are significant and for which it is not the case. It would be also interesting to check the influence of the meta parameters (K, size of the kernels) of the HMM model on the accuracy of our method.

In addition, in this paper, we only focused on the short-term memory data, an immediate extension of this work would be to test our methodology for data of the second session for the study of long-term image memorability. This methodology could be also combined to classical image memorability models based on image features to improve performance on individual scores. Another limitation is that HMM-based models only consider one previous time step. A future work could be to use deep learning methods for time series as RNN or LSTM. However, more data would be required to this end.

V. Conclusion

In this paper, an HMM based framework is proposed to analyze the gaze data as time series in the context of individual image memorability prediction. Results show significantly different eye patterns for correct and wrong responses, and the proposed scenarios for image prediction based on gaze data give promising results for the use of indirect measures in personalised QoE assessment. However, these results merit to be further investigate to improve performance and construct more concrete scenarios.

Acknowledgment

The work in this paper was funded from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 765911, European Training Network on Real Vision project.

REFERENCES

[1] Y. Rai and P. Le Callet, "Visual attention, visual salience, and perceived interest in multimedia applications," in *Academic Press Library in Signal Processing, Volume 6*. Elsevier, 2018, pp. 113–161.

[2] P. Le Callet and E. Niebur, "Visual attention and applications in multimedia technologies," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2058–2067, 2013.

[3] K.-c. Chen and H. J. Choi, "Visual attention and eye movements," 2008.

[4] Y. Rai, P. Le Callet, and G. Cheung, "Role of hevc coding artifacts on gaze prediction in interactive video streaming systems," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3344–3348.

[5] T. Chuk, A. B. Chan, and J. H. Hsiao, "Understanding eye movements in face recognition using hidden markov models," *Journal of vision*, vol. 14, no. 11, pp. 8–8, 2014.

[6] E. J. David, P. Lebranchu, M. P. Da Silva, and P. Le Callet, "Predicting artificial visual field losses: a gaze-based inference study," *Journal of Vision*, vol. 19, no. 14, pp. 22–22, 2019.

[7] A. Coutrot, J. H. Hsiao, and A. B. Chan, "Scanpath modeling and classification with hidden markov models," *Behavior research methods*, vol. 50, no. 1, pp. 362–379, 2018.

[8] C. Kanan, D. N. Bseiso, N. A. Ray, J. H. Hsiao, and G. W. Cottrell, "Humans have idiosyncratic and task-specific scanpaths for judging faces," *Vision research*, vol. 108, pp. 67–76, 2015.

[9] W. Poynter, M. Barber, J. Inman, and C. Wiggins, "Individuals exhibit idiosyncratic eye-movement behavior profiles across tasks," *Vision research*, vol. 89, pp. 32–38, 2013.

[10] P. Le Callet, S. Möller, and A. Perkis, "Qualinet White Paper on Definitions of Quality of Experience," COST IC 1003, Tech. Rep., 2012.

[11] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *CVPR 2011*. IEEE, 2011, pp. 145–152.

[12] Y. Baveye, R. Cohendet, M. Perreira Da Silva, and P. Le Callet, "Deep learning for image memorability prediction: The emotional bias," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 491–495.

[13] J. Fajtl, V. Argyriou, D. Monekosso, and P. Remagnino, "Amnet: Memorability estimation with attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6363–6372.

[14] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2390–2398.

[15] R. Cohendet, A.-L. Gilet, M. P. Da Silva, and P. Le Callet, "Using individual data to characterize emotional user experience and its memorability: Focus on gender factor," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–6.

[16] R. Cohendet, K. Yadati, N. Duong, and C.-h. Demarty, "Video Memorability To cite this version : HAL Id : hal-01785130 Annotating , Understanding , and Predicting Long-term Video Memorability," in *International Conference on Multimedia Retrieval*, 2018.

[17] C. A. McGrory and D. Titterington, "Variational bayesian analysis for hidden markov models," *Australian & New Zealand Journal of Statistics*, vol. 51, no. 2, pp. 227–244, 2009.

[18] E. Coviello, G. R. Lanckriet, and A. B. Chan, "The variational hierarchical em algorithm for clustering hidden markov models," in *Advances in neural information processing systems*, 2012, pp. 404–412.

[19] E. Coviello, A. B. Chan, and G. R. Lanckriet, "Clustering hidden markov models with variational hem," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 697–747, 2014.

[20] P. Lang, M. Bradley, and B. Cuthbert, "International affective picture system (iaps): affective ratings of pictures and instruction manual. university of florida, gainesville," Tech Rep A-8, Tech. Rep., 2008.