Communications in Information and Systems
Volume 00, Number 00, 1–20, 2015

Prediction of human looking behavior using interest-based image representations

Jong-Shenq Guo *, Karen Guo *, Paul Schrater

Looking behavior allows human to understand and interact with an enormous amount of information, a capacity challenging to replicate in AI systems. One of the core elements of this work is an effort to predict scan-paths from a combination of image information and past looking behavior. The success of this scan-path predication relies heavily on whether this image information can provide a sufficiently rich representation for prediction. In this paper, we show that changing representations dramatically simplifies and improves predictions of looking behavior. We introduce a representation of looking behavior that centers around interest-regions in images, defined by natural and collective looking behavior. These regions (called *interest-based regions*) can be used to partition images for semantic labeling and to provide a basis for shared representation across observers. Without any additional label or image information, we achieve highly accurate sequence prediction using this interest-based image representation.

1. Introduction

It is natural for humans to understand and interact with an enormous amount of information from the surrounding environment. But it is much more difficult for computers to process incoming information like humans. To make a computer more like a human, researchers in Artificial Intelligence (AI) give a direction to computers that simulate how a human understands the world, which could lead to the development of new computer technology. Specifically, the human visual system is interesting because it receives a huge amount of visual information through the eyes [17]. And human

[0]Corresponding author at: Department of Data Science and Big Data Analytics, Providence University, Taichung, Taiwan.
E-mail addresses: jsguo@mail.tku.edu.tw (J.-S. Guo), kguo2021@pu.edu.tw (K. Guo), schrater@umn.edu (P. Schrater)

visual attention is influenced by both this incoming visual information and other stimuli such as environment and each person's background knowledge. Humans analyze the incoming scenes over a series of looks, building up useful representations from outside information and generating high-quality interpretations to create a conscious understanding or reaction to the surrounding [39].

Yet human vision is influenced by incoming information remains concealed, which is critical for simulating this visual processing procedure by computers. One of the ways to connect visual information and human knowledge is object categorization. For example, ImageNet [12] is an image dataset that contains thousands of object classes and is used to train computers to detect and recognize these objects. And Alexnet [26] is one of the famous deep neural networks for retrieving information from the dataset and performing well on object detection and recognition tasks. However, humans consider not only the objects in the image but also the details or relations among them. The viewers will correspondingly change their eye movements based on where in the image they are interested in and which part of the image provides information that can benefit their interpretation. The difference in scanpath while giving the observer different tasks has been noticed by Yarbus back in 1967 [42]. Itti et al. [20] provided a model of generating *saliency map* by fusing intensity, color, and orientation. The so-called *saliency map* is a way to highlight which area is more important to the human or attracts more attention from human eyes within an image. Based on the procedure of generating the saliency maps [5], these attention models can be categorized as *bottom-up* [20, 18, 16] and *top-down* [23, 41]. They can also be stated as stimulus-driven and goal-driven, considering the process of how they solve their task. With more and more deep learning frameworks and models being established, most of the state-of-the-art saliency detection models include a neural network structure and use features learned from large-scale computer vision deep learning models [29, 27]. The saliency map also provides an efficient approximate solution to problems related to high-level visual concepts. For example, the concept of saliency map has been applied in [8, 14] as a prior probability distribution in object detection. A similar concept is used for finding the coexistence of specific objects in a set of images [9], video compression [19], foreground-background separation [35], and other visual tasks.

Saliency maps deliver a static and spatial distribution of visual attention, yet the human visual system (HVS) changes dynamically across time. Modeling visual attention temporally and spatially is much closer to the

mechanism of human cognitive behavior while viewing. There are characteristic allocation patterns over time, shared across people and driven by task domains. Based on the explanation on Wikipedia, *'In computer vision, a saliency map is an image that highlights the region on which people's eyes focus first.'* In addition, the saliency map is regularly colored to represent the importance of each pixel in a static image. While saliency map models being established for a couple of years, scanpath modeling has not received much attention until recent years. Scanpath is defined as a series of fixations and saccades on an image. Compared to the static saliency map, literature related to dynamic visual attention prediction is growing within the topic of scanpath prediction [28].

Scanpath prediction is a problem that focuses on developing methods for generating eye fixation sequences based on the given image. Starting from the early 2000s, Lee and Yu [30] introduced an information maximization framework during the targeted eye movements. Later, Wang et al. [40] proposed a computational model that predicts saccadic eye movement in natural images. The model generates sequential saliency maps based on integrating the responses from designed filters applied to each eye movement. The concept of representing the fixation points with the regional or local feature was used in Jiang et al. 's approach [21]. They used superpixel, which is a large unit of representing the element in an image defined in this paper, to indicate fixation points and applied least-squares policy iteration for learning. In addition to the better performance than the previous models, it is shown in [21] that the combination and comparison of several features show that the representation of the image and fixation point takes a significant place in solving the scanpath prediction problem. On the other hand, with the growth of deep learning models and the inspiration of human vision in neuroscience, neural network structures are applied to achieve both saliency and scanpath prediction. Kerkouri et al. [25] proposed an end-to-end deep-based model for predicting the scanpath based on the features of the saliency map simultaneously generated by their deep model. Assens et al. [3] proposed Saltinet with a CNN encoder-decoder network. They also came up with PathGan [2], a model for saliency prediction with a generative adversarial network (GAN) and LSTM layers. GAN has also been applied to predict realistic scanpath for the panorama images by Martin et al. [33]. The idea of using the recurrent neural network (RNN) structure also appeared in the model proposed by Chen et al. [10] and Sun et al. [38]. The ability of RNN and LSTM to process temporal relations between data makes them a critical role while predicting scanpath and strengthening the sequential dependency between fixations.

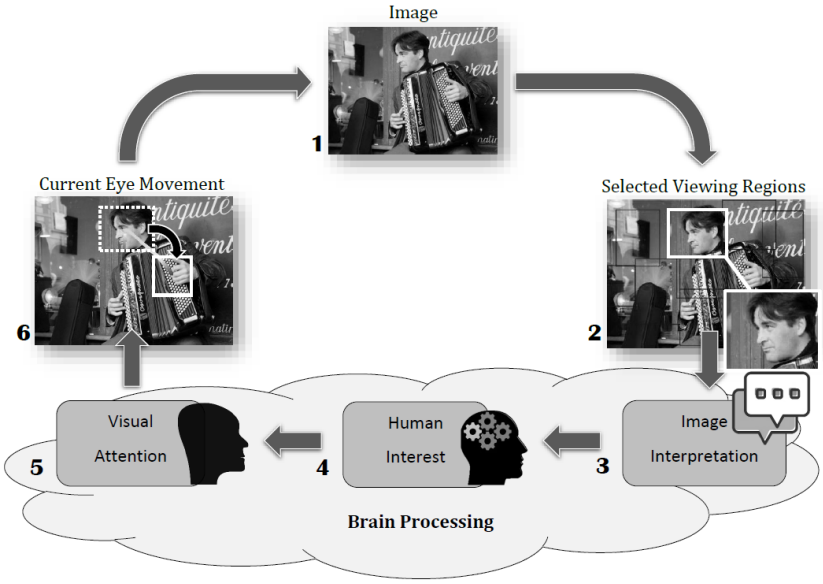

Figure 1: **Concept Map of how Humans interpret an Image.** Image understanding, human interest, visual attention and related looking behavior form a system of human viewing process that runs through time and iteratively updates information of each stage.

This paper proposes a new representation method that centers around interest regions in images. These regions, which we defined as *interest-based regions*, can be used to partition images for semantic labeling and provide a basis for shared representation across observers to predict their looking behavior. We first proposed a flow of how humans interpret an image with their visual attention that is detailed in Figure 1. Given an image, the information of the region that the person is viewing is obtained and interpreted by their brains (Stage 1 to 3). Then the region interpretation would affect their interest in this image for understanding more contents in the image (Stage 3 to 4). After deciding what is interesting after knowing this region, the visual attention would change together and affect the next eye movement (Stage 4 to 5 and 6 back to 1). We then formulate this looking behavior as a partial-observable Markov decision process (POMDP) (Figure 2). For recent studies which formulated the looking behavior as a Markov decision process, we refer the reader to, e.g., [36, 24, 11, 32]. We define a state space as a belief

vector over the set of interest regions, represented by occupancy distributions over the image. It is shown that using an interest-based image representation makes the scan-path prediction problem more feasible. We achieve highly accurate sequence prediction without any additional label or image information. These results show that observers use highly structured and consistent strategies while free-viewing images. This consistency has been masked by the use of feature-based and object-based image representations that do not capture the interpretative structure underlying human looking. Based on their ability to accurately capture looking behavior, interest-based models can dramatically improve applications that can capitalize on looking behavior, including image captioning, advertisements, and diagnostics. Specifically, some further details of an application on diagnostics is to be given in the discussion session.

2. Modeling Human Looking Behavior

Many scanpath prediction methods focus on learning the point-to-point pattern and have a decent prediction accuracy. However, human eyes usually see the surrounding region of the fixated location. Therefore, it is important to learn the transition between regions instead of points. Here we start with formulating this looking behavior as a decision process on interpreting an image and introducing our new representation of eye fixation as a belief state over a set of interest regions.

2.1. Partially Observable Markov Decision Process (POMDP)

The POMDP was first described by Karl Johan Åström in 1965 [4]. It is normally defined by the tuple $(S, A, T, V, Z, O, \gamma)$, where

- S is a set of *latent* states s,
- A is a set of actions a,
- T is the transition probability function $T(s'|s, a)$,
- $V : S \times A \to \mathbb{R}$ is the value (reward) function,
- Z is a set of observations z,
- O is the observation probability function $O(z|s', a)$,

- $\gamma \in [0, 1]$ is the discount factor.

While the environment is in a state $s \in S$, the agent takes an action $a \in A$, transitioning the state from s to s' with probability $T(s'|s, a)$ at each time step. The agent receives an observation $z \in Z$, related to the state s and the action a by the observation probability $O(z|s, a)$. In addition, the agent also receives a reward signal $v = V(s, a)$ and then this process repeats.

Since POMDP contains latent states S that cannot be directly observed, the agent must makes its decisions with the uncertainty of S. Therefore, it is common to formulate a POMDP as a *belief MDP* with a new-defined observable *belief*. *Belief* b is defined as a probability distribution over the state space S, and $b(s)$ denotes the probability that the environment is in state s. Following this state transition method, belief MDP can thus be defined as a tuple (B, A, f, Ψ, γ), where

- B is a set of *non-latent* beliefs $b(s)$ of states s,
- A is a set of actions a,
- f is a belief state transition function: $b' = f(b, a, z)$; $f \propto O(z|s', a) \sum_{s \in S} T(s'|s, a)b(s)$,
- $\Psi : B \times A \to \mathbb{R}$ is the reward function on belief states: $\psi = \Psi(b, a) = \sum_{s \in S} V(s, a)b(s)$,
- $\gamma \in [0, 1]$ is the discount factor the same as the original POMDP.

Then the original POMDP is reformulated into a MDP with observable belief states B. This is the decision process to be used in this work.

Variable	Descriptions
x_t	a set of fixation points: proxy for a_t
r_t	inferred region of fixation x_t based on interest regions $\{k\}$
d_t	(local) interpretation of r_t belief vector over interest regions $\{k\}$
s_t	(global) current understanding of the given image based on former beliefs
a_t	movement from the previous fixation point to the next one

Table 1: Variables that are related to viewing process and are updated across time.

With the above definition, we can then connect some notations in MDP with fixations and interpretations on an image in table 1. In addition, by the definition in table 1, we reformulate the procedure described in Figure 1 and describe the whole iterated image interpretation process of human looking behavior in Figure 2 with details as following:

Given an image I, the subject was asked to view the image freely, which means the only goal is to understand the content without answering any task-oriented questions. While starting viewing the image, the subject may have their eyes wandering around the image to gain more knowledge of this image. Let their current fixation point be x_t. The interpretation surrounding the observation x_t is denoted as d_t, which is not observable. This local interpretation around x_t provides knowledge for the further understanding of the whole image. Let the current understanding of the image be s_t, and this would then be the (latent) knowledge state that is updating while a new x_t is observed.

In addition to the directly observable fixation sequences $\{x_t\}$, the observer's fixations in an image also give us indirect information about their local interpretation. We treat the problem of estimating the local interpretation as having two parts. First we infer which annotations are relevant by a probabilistic weighting of the annotated regions intersected by fixation x_t. Secondly, we marginalize the latent interpretation state to focus on prediction of subsequent belief probability and fixations. By marginalizing the dependency between b_t and other variables, we preserve only observable variables and simplify it to be a temporal-series data analysis problem.

In the next subsection, we introduce **interest-based region** as the basis for belief vector of local interpretation of an image.

2.2. Interest-based Regions Representation

In this section, we introduce interest-based region representation to provide a way to represent the common regions that humans tend to be more interested in while viewing. Interest-based region (IbR) representation is based on the collected human eye fixations on images. In addition, IbR can be explicitly represented as a probability distribution, which makes an objective way to describe the subjective idea. Our approach requires human visual attention data, yet collecting large-scale attention data is always the main reason for prohibiting the analysis of different aspects and attributions. Huang et al. introduced an approximation method *SALICON* to visual attention via the mouse trace [22] and a crowdsourcing platform to collect large-scale

attention data. They first applied a Gaussian blur filter on the images from the MS COCO dataset [31]. Then these images were uploaded to the platform of Amazon Mechanical Turk to collect large-scale mouse-tracking data. The collected mouse traces on the blurred images can be transformed into simulated eye movement maps on the images. In this way, the visual attention map of an image can be approximated from a large amount of subjects' mouse traces instead of using an eye-tracking machine. From their analysis, these mouse-tracking resulting attention maps can simulate the same or even better attention map compared to the one simulated by other saliency detection algorithms. And thus, the mouse traces can approximate eye movements and are sufficient to represent human visual attention.

Here in our approach, we first recorded 104 subjects' mouse traces on given images with SALICON [22] to simulate their eye-movements. To emphasize the points that attract people to fix their attention, we came up with a way to define "fixation points" from the collected mouse traces. The fixation points are defined and filtered by the length of time that the mouse stays at a certain position. We then apply a Gaussian blur filter to fuse all fixation points in an image and generate the overall visual attention map, or so-called saliency map. From these saliency maps, we can see that there are always some brighter regions, which means these regions include significant parts of fixation points inside and have certain contents or information that attract most people's visual attention. However, the definition of these "attention-focused" regions remains unclear. Therefore, we propose a clustering method to generate these informative regions explicitly as follows.

Our method starts with the input of overall fixation points $\{x^j\}$ of an image j. With the generated saliency map M^j, our goal is to find the most outstanding region centers to separate the fixation points into meaningful regions. We first use a sliding square window with given width ρ to search whether the center of this window has the maximum value in M^j. If so, this point is saved as a center reference $\{r_k^j\}$ and the window keeps moving on to the search for the next local maximum value until the window hits the bounds. With these centers $\{r_k^j\}$ being found, we assign each fixation point x to a certain region by finding the index k with the shortest distance $||x - r_k^j||$ among all possible centers. The fixation point x is then denoted by x_k^j. Then the variables of Gaussian mixture model G^j: mean μ_k^j and covariance σ_k^j can be calculated from all fixation points $\{x_k^j\}$ assigned to each region k of image j. We call the resulting regions ***interest-based regions***, because these regions are not only simply just the clustering results of fixation points, but also the parts that most subjects decide to focus on while viewing the image and understanding the contents inside. With these regions

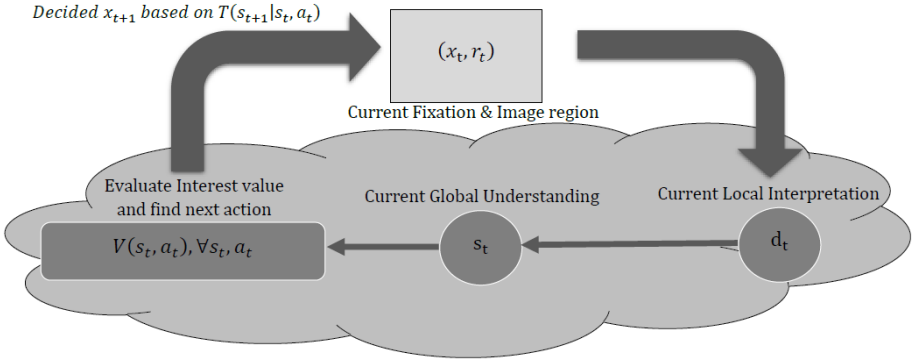

Figure 2: **Formalization of Image Interpretation as a Partially Observed Decision Process**. Graphical schematic of the structure of the optimal policy for image interpretation. We will use this structure to focus our policy learning methods.

and the model G^j being generated, we separate all the fixation points into distinct regions that include major amount of visual attention and meaningful contents. These regions are contextually informative, which has been used and discussed in [15].

2.3. Gaussian Mixture Posterior as Belief Vector

By having the Gaussian mixture model G^j for each image j, we are able to assign each fixation point to the closest region. However, instead of having the one-hot vector as a representation of a fixation point, we would like to consider a representation method with a continuous property that allows derivative and reconstruction of the original point. Therefore, we use the posterior function generated from the normalization between the Gaussian mixture model G^j and the uniform distribution as background. Let $L_k^j(x)$ be the likelihood of region k and point x from the model G^j in image j. We combine L_k^j with the uniform distribution and generate a vector-valued function $H(x)$ as following:

$$H(x) = [L_1^j(x), L_2^j(x), \cdots, L_K^j(x), \frac{1}{\#pixels}], x \in I^j, \tag{2.1}$$

where K is the number of interest-based regions in image j. Then the posterior can be written as:

$$P(x) = [P_1(x), \cdots, P_K(x), b],\ P_k(x) = \frac{L_k^j(x)}{\hat{H}(x)}, \tag{2.2}$$

where b, the value of normalized background uniform distribution, and $\hat{H}(x)$ are defined by

$$b = \frac{1}{\hat{H}(x)(\#pixels)},\ \hat{H}(x) = \sum_{k=1}^{K} L_k^j(x) + \frac{1}{\#pixels}.$$

This way, we can reconstruct the fixation point x from the posterior $P_k(x)$ by giving the posterior maps T_k^j for each group k generated from each component in G^j. The reconstruction y of a fixation point x based on the posterior can be formulated as the following:

$$y = \mathrm{argmin_y} \prod_{\mathrm{k=1}}^{\mathrm{K}} (\exp \mathrm{P_k(x)} - \mathrm{T_k^j(y)})^2. \tag{2.3}$$

With the posterior as a representation of each fixation point, we can flexibly include the relation between these interest-based regions and map the representation as a belief vector over the interest-based regions. This way, the posterior not only provides an informative representation way of fixation points, but also generates the belief state that can be formulated in the belief MDP described above.

3. Prediction of Human Looking Behavior: Experiments and Results

In this section, we shall describe our experimental setting and the results we obtained for the prediction problem of human looking behavior.

3.1. Experimental Setting

Our stimuli images come from the dataset generated by the clinical psychology department [15]. Images of this dataset are collected from both IAPS

(International Affective Picture System) [7] and other public domain websites with a variety of contents and backgrounds such as face-included or face-excluded images and clutter or clean background. To generate interest-based regions and their posterior distributions, we use the eye movement data collected from 104 subjects by SALICON [22].

We have shown that our image interpretation problem can be formulated as a MDP model in the previous section. Furthermore, by marginalizing the states that are not observable, we can treat the MDP model as a temporal-series data analysis problem (Figure 3). Here we use a stack of Long Short Term Memory (LSTM) to predict the next state. We use the historic set of posterior probability $\{P(x_i), i = 1, \cdots, w\}$ as the input state of LSTM for predicting the next state $P(x_{w+1})$, where w is the length of look-back posterior. Moreover, in order to evaluate the performance of our interest-based regions representation, we apply the same temporal model to object-based representation, which generated from the LabelMe object segmentation [37] of each image. On the other hand, to be more precise on evaluation, we considered the evaluation method based on the MIT1003 and SALICON datasets. MIT1003 is a dataset that contains 1003 images of natural indoor and outdoor scenes, along with the eye-tracking data collected from 14 observers. This dataset has been used as a benchmark for saliency prediction [21]. Yet the temporal relation is kept in the dataset, which makes it a good benchmark for scanpath prediction. Due to the limitation of computation power of our system, we choose 200 images from the dataset and run the training process on the data of 01 observers. The prediction of our trained model is applied to the rest 4 observers. In addition, we also consider the SALICON dataset, which contains the mouse traces on 10000 images for training and 5000 images for validation. Since each image has different numbers of observers, here we use 70 percent observers for training and the rest 30 percent for testing.

3.2. Results

With the development of scanpath prediction modeling, a few scanpath comparison metrics are established for comparing the predicted scanpath with the ground truth [1]. Yet few evaluation metrics for scanpath prediction are fitted to *region-based scanpath*. Here we use accuracy for evaluating the region-match rates of interest-based regions prediction. In addition, we also

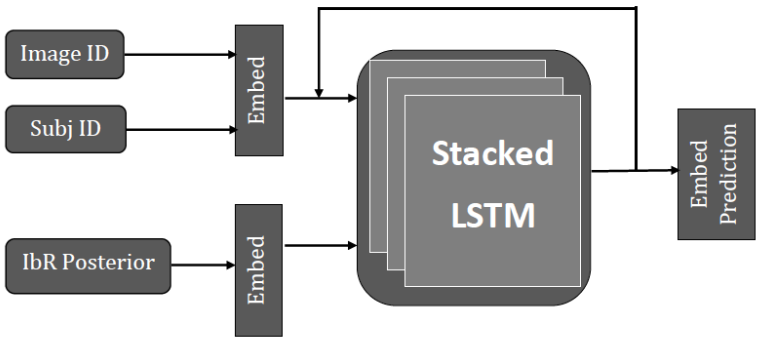

Figure 3: Based on the temporal relation between belief states b and the action a, the structure of Markov Decision Process here can be simplified to fit any Recurrent Neural Network type of deep learning structure. In this figure, IbR stands for Interest-based Region and LSTM is the Long Short Term Memory.

generated a scanpath by the predicted posterior based on our Gaussian mixture models. This way, we could apply scanpath comparison metrics to compare our results with other methods. Here we consider MultiMatch method [13] to be the evaluation metric for scanpath prediction. MultiMatch has five measures: shape, direction, length, position, and duration, to indicate the similarity between the predicted scanpath and ground truth. Starting with the simplification step, MultiMatch temporally aligned two simplified scanpaths so that the following comparison by each metric is well-defined. MultiMatch Shape considers the vector difference, MultiMatch Length considers the length difference between the endpoints of two saccade vectors, MultiMatch Direction measures the angular distance between saccade vectors, and MultiMatch Position finds the difference in position between aligned fixations. MultiMatch Duration considers the difference in fixation duration between aligned fixations. Since our method does not generate a prediction on the duration, we only consider the first four measurements of MultiMatch. For object-based representation, we apply multiple metrics described in [6]: Exact Match Ratio, Precision, Recall, and F1.

We also test on whether the number of previous historical attentions would affect the prediction. In Table 2, we see that the evaluation result is really bad without including any previous attention. However, by including simply one previous attention it exhibits a significant improvement, and

Metric \Historic Data Length	1	2	10
Accuracy (Int)	0.1393	**0.9184**	**0.9269**
Exact Match Ratio (obj)	0.2283	0.7824	0.7992
Precision (obj)	0.2079	0.8166	0.8236
Recall (obj)	**0.2678**	0.8151	0.8264
F1 (obj)	0.2097	0.8117	0.8215

Table 2: Comparison of evaluation metrics of Interest-based Regions and Object-based Regions considering different historic data length.

Model	MM Shape	MM Dir	MM Len	MM Pos
PathGan[2]	0.9608	0.5698	0.9530	0.8172
Le Meur[34]	0.9505	0.6231	0.9488	0.8675
SALYPATH[25]	0.9659	0.6275	0.9521	0.8965
bR + SALICON	0.9976	0.6411	0.9964	**0.9687**
IbR + MIT1003	**0.9993**	**0.8245**	**0.9990**	0.9343

Table 3: Overview of the MM (MultiMatch) measurements on different scanpath prediction methods with SALICON and MIT1003 dataset and as a reference comparing to our method. IbR stands for Interest-based Region

including more history attentions only provide a slightly improvement on performance. Our prediction based on interest-based regions outperforms any approach measurement of object-based regions. This also shows that our interest-based regions contain meaningful and critical information for prediction.

In Table 3, we provide a referencing numbers of MM measurements on SALICON dataset with several state-of-the-art scanpath prediction methods as an objective reference of how our prediction performs. The MM measurements are also applied to our predicted scanpath from our dataset. From the results above, we can see that our model not only predict the next interest-based region accurately based on historical viewing, the model also provide a good performance on scanpath prediction based on our Gaussian Mixture Model posterior calculation.

4. Discussion

In the experiment session, we received a good performance in predicting the next fixated region by only using interest-based region representation. With the property of switching attention from region to region, the probability distribution that describes the current location, either a region or a specific point, becomes essential considered the simulation of this information delivering process. Most of the representation methods require an additional simulation of the probability that describes the representation. Consider this situation, the Gaussian mixture posterior mentioned in section 2.3 related to the interest-based region not only provides a way to connect with the decision process as a belief vector. The Gaussian mixture posterior provides a probability distribution that well-described the interest-based region representation. This way, the information in our representation will not be reduced during the addition simulation step.

Our LSTM learning framework provides a flexible structure for adding more information, such as additional visual or descriptive features, with interest-based region representation while training. In addition, our prediction result also provides a potential performance increase by incorporating more information mentioned above.

Finally, we briefly provide a description of a real-world application on diagnostic based on the interest-based region representation. We have conducted a study on objective psychological symptom diagnosis. In this study, a probability classifier is constructed based on the interest-based image representation to differentiate between *bipolar, schizophrenia, and control groups*. The performance of the classification shows that involving interest-based regions representation as a hidden user state forms a feasible user model between humans' eye movement and psychological symptoms. With the usage of interest-based region representation, the overall system described in this study provides a noninvasive method for providing numerical measurement and prediction of groups. More precisely, we have collected fixation data from three groups of people: bipolar, schizophrenia, and health-control groups viewing the 113 images, which is the same image dataset mentioned earlier in section 3.1 [15]. The wide variety of these selected images also has the potential to be a critical factor in diagnosing bipolar, schizophrenia, or other psychological symptoms. This is a preliminary sample set since the data collection from each group of specific psychological symptoms is extremely difficult. The health-control group contains 5 subjects, whereas bipolar and schizophrenia have 11 subjects each. Due to the lack of data, we only consider validating our method at this stage to prove the feasibility

of recognizing subjects from different groups with sufficient data. The pilot result indicates that our system has the power to find the variation between groups and apply it as the critical component to distinguish different groups. This also indicates the potential applicability of the interest-based region representation method. We expect that, with more data being involved in learning the weights and training the classifier, fewer subjects will be left out, and prediction will be substantially more accessible with less need for data collection (fewer test images).

Acknowledgements

This work was supported in part by the Ministry of Science and Technology of Taiwan under the grants 111-2115-M-032-005 (JSG) and 111-2115-M-126-001 (KG). We would like to thank the anonymous referee for some valuable comments. Also, appreciation on the help from Ching-Yao Chen, Yu-Fang Chou, and Chiao-Ying Cheng from Provident University, who provide helps on running and evaluation of MultiMatch on SALICON and MIT1003.

References

[1] Nicola C. Anderson, Fraser Anderson, Alan Kingstone, and Walter F. Bischof. "A comparison of scanpath comparison methods". In: *Behavior research methods* 47 (2015), pp. 1377–1392.

[2] Marc Assens, Xavier Giro-i-Nieto, Kevin McGuinness, and Noel E. O'Connor. "PathGAN: visual scanpath prediction with generative adversarial networks". In: *European Conference of Computer Vision* (2018).

[3] Marc Assens, Xavier Giro-i-Nieto, Kevin McGuinness, and Noel E. O'Connor. "Saltinet: Scan-path prediction on 360 degree images using saliency volumes". In: *ICCV* (2017).

[4] Karl Johan Åström. "Optimal control of Markov processes with incomplete state information". In: *Journal of Mathematical Analysis and Applications* 10 (Feb. 1965), pp. 174–205. DOI: 10.1016/0022-247X(65)90154-X.

[5] Ali Borji and Laurent Itti. "State-of-the-art in visual attention modeling". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (2013).

[6] Ali Borji, Dicky N. Sihite, and Laurent Itti. "What stands out in a scene? A study of human explicit saliency judgment". In: *Vision Research* 91 (2013). DOI: 10.1016/j.visres.2013.07.016.

[7] Margaret M. Bradley and Peter J. Lang. *International Affective Picture System (IAPS)*. 2015. URL: http://csea.phhp.ufl.edu/media/iapsmessage.html.

[8] Kai-Yueh Chang, Tyng-Luh Liu, Hwann-Tzong Chen, and Shang-Hong Lai. "Fusing generic objectness and visual saliency for salient object detection". In: 2011, pp. 914–921.

[9] Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai. "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model". In: *CVPR* (2011), pp. 2129–2136.

[10] Zhenzhong Chen and Wanjie Sun. "Scanpath prediction for visual attention using IOR-ROI LSTM". In: *International Joint Conference of Artificial Intelligence* (2018), pp. 642–648.

[11] Antoine Coutrot, Janet H. Hsiao, and Antoni B. Chan. "Scanpath modeling and classification with hidden Markov models". In: *Behavior Research Methods* 50 (2018). DOI: 10.3758/s13428-017-0876-8.

[12] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 2–9. DOI: 10.1109/CVPR.2009.5206848.

[13] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. "It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach." In: *Behavior research methods* 44 (2012), pp. 1079–1100.

[14] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. "Context-aware saliency detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (10 2012), pp. 1915–1926. DOI: 10.1109/TPAMI.2011.272.

[15] Karen Guo, Danielle N Pratt, Angus Macdonald, and Paul R Schrater. "Labeling images by interpretation from Natural Viewing". In: *CEUR Workshop Proceedings* 2068 (2018).

[16] Jonathan Harel, Christof Koch, and Pietro Perona. "Graph-based visual saliency". In: 2006, pp. 545–552. DOI: 10.1.1.70.2254.

[17] Helene Intraub. "The representation of visual scenes". In: *Trends in cognitive sciences* 1 (1997), pp. 217–222.

[18] Laurent Itti. "Models of bottom-up attention and saliency". In: *Neurobiology of Attention* (2005), pp. 576–582. DOI: 10.1016/B978-012375731-9/50098-7.

[19] Laurent Itti. "Realistic avatar eye and head animation using a neurobiological model of visual attention". In: *Proceedings of SPIE* 5200 (2004), pp. 64–78.

[20] Laurent Itti, Christof Koch, and Ernst Niebur. "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (1998), pp. 1254–1259.

[21] Ming Jiang, Xavier Boix, Gemma Roig, Juan Xu, Luc Van Gool, and Qi Zhao. "Learning to Predict Sequences of Human Visual Fixations". In: *IEEE Transactions on Neural Networks and Learning Systems* 27 (2016), pp. 1241–1252.

[22] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. "SALICON: Saliency in Context". In: vol. 07-12-June. 2015, pp. 1072–1080. ISBN: 9781467369640. DOI: 10.1109/CVPR.2015.7298710.

[23] Tilke Judd, Krista A Ehinger, Frédo Durand, and Antonio Torralba. "Learning to predict where humans look". In: *ICCV* (2009), pp. 2106–2113.

[24] Adam B. Kashlak, Eoin Devane, Helge Dietert, and Henry Jackson. "Markov models for ocular fixation locations in the presence and absence of colour". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67 (2018), pp. 201–215.

[25] Mohamed A. Kerkouri, Marouane Tliba, Aladine Chetouani, and Rachid Harba. "SALYPATH: A deep-based architecture for visual attention prediction". In: *Proceedings - International Conference on Image Processing, ICIP* 2021-September (2021), pp. 1464–1468.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* (2012), pp. 1097–1105.

[27] Matthias Kummerer, Thomas S.A. Wallis, Leon A. Gatys, and Matthias Bethge. "Understanding Low- and High-Level Contributions to Fixation Prediction". In: Institute of Electrical and Electronics Engineers Inc., Dec. 2017, pp. 4799–4808.

[28] Matthias Kümmerer and Matthias Bethge. "State-of-the-art in human scanpath prediction kummerer". In: *arXiv preprint* (2021).

[29] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. "Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet". In: *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings* (2015).

[30] Tai Sing Lee and Stella Yu. "An Information-Theoretic Framework for Understanding Saccadic Eye Movements". In: *Neural Information Processing Systems* (2000).

[31] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft COCO: Common objects in context". In: 2014, pp. 740–755.

[32] Noa Malem-Shinitski, Manfred Opper, Sebastian Reich, Lisa Schwetlick, Stefan A. Seelig, and Ralf Engbert. "A mathematical model of local and global attention in natural scene viewing". In: *PLoS Computational Biology* 16 (2020).

[33] Daniel Martin, Ana Serrano, Alexander W. Bergman, Gordon Wetzstein, and Belen Masia. "ScanGAN360: A Generative Model of Realistic Scanpaths for 360° Images". In: *IEEE Transactions on Visualization and Computer Graphics* 28 (2022), pp. 2003–2013.

[34] Olivier Le Meur and Zhi Liu. "Saccadic model of eye movements for free-viewing condition". In: *Vision Research* (2015). ISSN: 18785646. DOI: 10.1016/j.visres.2014.12.026.

[35] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. "Saliency Filters: Contrast Based Filtering for Salient Region Detection". In: 2012.

[36] Yashas Rai, Patrick Le Callet, and Gene Cheung. "Qualitifying the relation between perceived interest and visual salience during free viewing using trellis based optimization". In: *IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)* (2016), pp. 1–5.

[37] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. "LabelMe: A database and web-based tool for image annotation". In: *International Journal of Computer Vision* 77 (2008), pp. 157–173.

[38] Wanjie Sun, Zhenzhong Chen, and Feng Wu. "Visual scanpath prediction using IOR-ROI recurrent mixture density network". In: *IEEE transactions on pattern analysis and machine intelligence* 43.6 (2019), pp. 2101–2118.

[39] Anne M. Treisman and Garry Gelade. "A feature-integration theory of attention". In: *Cognitive Psychology* 12 (1980), pp. 97–136.

[40] Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. "Simulating human saccadic scanpaths on natural images". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2011), pp. 441–448.

[41] Jimei Yang and Ming-Hsuan Yang. "Top-Down Visual Saliency via Joint CRF and Dictionary Learning". In: *IEEE transactions on pattern analysis and machine intelligence* (2012).

[42] Alfred L. Yarbus. *Eye movements and vision*. Springer, 1967.