Appears in
The Annual Meeting of The Cognitive Science Society (CogSci), Berlin, 2013

Understanding eye movements in face recognition with hidden Markov model

Tim Chuk (u3002534@connect.hku.hk)[1]
Alvin C. W. Ng (asangfai@gmail.com)[2]
Emanuele Coviello (ecoviell@ucsd.edu)[3]
Antoni B. Chan (abchan@cityu.edu.hk)[2]
Janet H. Hsiao (jhsiao@hku.hk)[1]

[1] Department of Psychology, The University of Hong Kong, Pokfulam Road, Hong Kong
[2] Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[3] Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA, USA

Abstract

In this paper we propose a hidden Markov model (HMM)-based method to analyze eye movement data. We conducted a simple face recognition task and recorded eye movements and performance of the participants. We used a variational Bayesian framework for Gaussian mixture models to estimate the distribution of fixation locations and modeled the fixation and transition data using HMMs. We showed that using HMMs, we can describe individuals' eye movement strategies with both fixation locations and transition probabilities. By clustering these HMMs, we found that the strategies can be categorized into two subgroups; one was more holistic and the other was more analytical. Furthermore, we found that correct and wrong recognitions were associated with distinctive eye movement strategies. The difference between these strategies lied in their transition probabilities.

Keywords: Hidden Markov Model (HMM); eye movement; scan path; holistic processing; face recognition.

Introduction

In the late 19th century, soon after Edmund Huey's invention of the world's first eye tracker, researchers discovered that in many daily life activities, eye movements were rapid, discontinuous, and interrupted by temporary fixations (Wade & Tatler, 2011). Nowadays, this finding has been widely accepted and described as the 'saccade and fixate' strategy (Land, 2011). Eye movements were found to facilitate face learning and recognition. For instance, Henderson et al. (2005) showed that when participants were restricted to view face images only at the center of the images, their recognition performances were significantly lowered than when they were allowed to view the images freely. Autistic patients, who could not judge facial expressions correctly, were found to have abnormal eye fixations patterns (Pelphrey et al, 2002).

Empirical studies on the relationship between eye movement and face recognition have primarily been focusing on identifying the regions of interest (ROIs). A ROI is a region on the face which people frequently fixate in, such as the two eyes. Early studies often divided a face into several regions and then identified the ROI through comparing the frequencies of each region being fixated in. However, this approach suffered from the lack of an objective manner to divide faces. For instance, Barton et al. (2006) defined the two eyes as two irregularly shaped ROIs, while Henderson et al. (2005) defined the two eyes as one ROI. Another problem is that the predefined ROIs may not really represent the data because different individuals have different saccade patterns. More recent studies attempted to discover ROIs directly from data. A commonly adopted way was to generate statistical fixation maps. A fixation map can be created by identifying the location of fixations and convolving a Gaussian kernel on each fixation. Two fixation maps can be compared by Pixel test, which discovers statistically significant differences in pixels (Caldara & Miellet, 2011). Using fixation maps, it was found that the upper center (i.e. the nose) and the upper left (i.e. the left half of the nose and the left eye) parts of a face were the two most frequently viewed areas (Hsiao & Cottrell, 2008). This result was consistent with an earlier study which used the Bubbles technique in discovering regions with diagnostic features in face recognition (Gosselin & Schyns, 2001). Fixation maps also showed that children from different cultural backgrounds demonstrated different eye fixation patterns (Kelly et al, 2011).

The use of fixation maps in face recognition studies had been fruitful. However, as discussed earlier, eye movements combine saccades and fixations. The fixations recorded in eye movement studies should be considered as *time-series data* that are collected over time. The eyes fixate at a location shortly, before a saccade brings them to the next location. Many studies showed that saccades can be influenced by top-down expectations as well as bottom-up inputs. Yarbus's (1965) well-known eye movement studies showed that depending on what people expect to see, they exhibited different saccade patterns when looking at the same target image. Mannan et al. (1997) discovered that saccades were more likely to be driven to the more 'informative' areas of an image, such as the edges and the high-spatial-frequency areas. These findings imply that the target location of a saccade could be a variable that has a set of possible values; different values could be associated with different probabilities. In this sense, eye movements may be considered as a stochastic process, which could be better understood using time-series probabilistic models. The fixation maps, however, do not contain temporal information.

Currently, there are two methods for describing the temporal information in eye movement data. One is the string-editing method. It requires an image to be divided into several ROIs, each labeled with a letter, so that a sequence of eye fixations can be described by a string. Two strings are then compared by measuring their Levenshtein distance (Goldberg & Helfman, 2010). This method does not capture the temporal information very precisely because the measure of Levenshtein distance does not precisely represent the sequential differences between two strings. For instance, the strings CAT and BAT differ in their first element, while the strings CAB and CAT differ in their last element. In both cases, however, the Levenshtein distance is one. The other method is to generate fixation maps by fixation and compare between conditions (Caldara & Miellet, 2011). For instance, if an experiment has two conditions, all the first fixations in each condition can be used to generate a fixation map. A comparison between the two fixation maps will show whether the two groups differ significantly in their first fixations. However, the problem associated with this method is that the significant areas are likely to be scattered so that the pattern could be hard to interpret. In this paper, we propose to use a time-series statistical model, the hidden Markov model (HMM) with Gaussian emission densities, to analyze eye movement data. We show that HMMs can 1) summarize a person's general eye movement strategy, including person-specific ROIs and saccade patterns, 2) reveal between-subject similarities and differences of eye movement patterns, and 3) discover the association between recognition performance and eye movement strategies. In the next section, we will 1) briefly describe the experiment in which we collected the data, and 2) explain the HMM-based method in more length.

Method

Experiment

A total of 32 Chinese participants were recruited at the University of Hong Kong. The experiment was divided into a training phase and a testing phase. In the training phase, participants were shown a total of 20 frontal face images. In the testing phase, participants were shown 40 frontal face images; 20 were new images and 20 were the ones appearing in the training phase. They were asked to judge whether they had seen the faces before. Their responses in the testing phase were recorded together with the fixations they made before the response. Eye movements were tracked and recorded using the Eyelink II eye-tracking system. On average, participants made 2.5 fixations per trial, ranged from one to three (this average was 1.8 fixations in Hsiao & Cottrell, 2008).

Model

HMMs are widely used to model data generated from Markov processes (Barber, 2012). A Markov process is a process whose present state is determined only by its previous state. The states in an HMM are not directly observable, so that the current state of the process can only be inferred from 1) the association between the assumed hidden state and the observed data, and 2) the likelihood of transiting to the assumed state from the previous state. The association among the observable data and the hidden states are summarized using probability distributions; each distribution represents the likelihood of a hidden state generating the data. The probabilities of transiting from one state to other states are summarized in a transition matrix; each element represents the probability of that transition. An HMM also has a vector of prior values; each value indicates the probability of the HMM starting from the corresponding state.

For instance, natural language processing is one area in which HMM has been widely applied. The observable data are the words in a corpus, and the hidden states are the word-class tags, such as nouns, verbs, and adjectives. An HMM cannot directly observe the word-class tags of the words, but can infer them from the observed words and the likelihood of transiting from one word-class to another.

In the context of face recognition, the HMM contains a number of hidden states, which each represents a different ROI of the face. The directly observable data is the fixation location, which belongs to a particular hidden state (ROI). The distribution of fixations in each ROI is modeled as a two-dimensional Gaussian distribution in a Cartesian space. Over time, the transition from the current hidden state to the next state represents the saccade pattern, i.e., movement between ROIs, which is modeled by the transition matrix of the HMM. In summary, the hidden states of the HMM correspond to the ROIs of the face, where each is observable through a two-dimensional Gaussian emission density of fixations, and the transitions between hidden states represent the saccade patterns.

Given a set of chains of fixations, we estimated the parameters of the HMM using a two-stage procedure. We first learned the ROIs on the face from the fixation data. Ignoring the temporal information, the ROIs can be seen as a mixture of two-dimensional Gaussian distributions, i.e., a Gaussian mixture model (GMM). In this study, we used the variational Bayesian framework for Gaussian mixture models (VBGMM) to estimate the Gaussian parameters, as well as the number of GMM components (Bishop, 2006). This Bayesian hierarchical method puts prior distributions on the GMM parameters, and uses approximation methods to find the *maximum a posteriori* (MAP) estimate. One important feature of VBGMM is that it can automatically estimate the optimal number of ROIs and 'deactivate' the redundant ones. After discovering the GMM components, or the ROIs, we next estimated the transition probabilities and prior probabilities of the hidden states, using the forward-backward algorithm (Bishop, 2006).

In this study, we aim to use HMMs to address two questions. Firstly, we wanted to discover the eye movement strategy of each individual in order to reveal the common strategies shared by a subgroup of the participants. Secondly, we wanted to explore whether accuracy at face recognition was related to eye movements. To address the first question, we trained one HMM per subject, using fixations collected from all the trials of the subject, in order to represent the general eye movement pattern of that sent the general eye movement pattern of that subject. To cluster the subjects' HMMs, we used the variational hierarchical EM algorithm (VHEM) for HMMs (Coviello et al, 2012). The VHEM algorithm takes HMMs as inputs, separates the inputs into subgroups, and estimates a representation HMM for each subgroup.

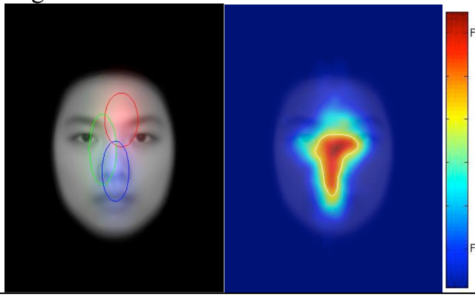
To address the second question, we trained two HMMs per subject, using fixation sequences collected from all the correct trials (i.e., correct HMM) and all the wrong trials (i.e., wrong HMM), respectively, to represent two eye movement strategies that led to different performances. We compared the correct HMMs to the wrong HMMs using subject analysis, based on the differences in log-likelihoods of the observed data, in order to examine whether eye movement strategies that lead to correct or wrong responses have significantly different patterns. Specifically, for the fixation sequences of a participant leading to correct responses, we calculated the log-likelihoods of observing the sequences from the correct HMM, and then computed the mean. We also calculated the mean log-likelihood from the wrong HMM using the same sequences. Doing this on all the 32 participants yielded two vectors of mean log-likelihoods, one represented the mean log-likelihoods of the correct HMMs generating the correct eye movements, and one represented the mean log-likelihoods of the wrong HMMs generating the correct eye movements. The differences between the mean log-likelihoods for each subject is an approximation to the Kullback-Leibler (KL) divergence between the correct HMM and the wrong HMM, which is a measure of difference between two distributions (Bishop, 2006). Similarly, we also calculate the mean log-likelihoods of the fixation sequences leading to incorrect responses under the wrong and correct HMMs.

Results

Section 1.1- Summary of all eye movement patterns

In order to model a participant's eye movement patterns, we pooled all the fixations that a participant made, regardless of their sequential order, and applied the VBGMM to discover a mixture of Gaussian distributions. We then used the found Gaussian components and the fixations in the forward-backward algorithm to estimate the transition probabilities and the prior values of the Gaussian components. The fixations put into the forward-backward algorithm were in their sequential orders. Each participant's eye movements were modeled by an HMM. Using the VHEM to group all HMMs into one cluster, the VHEM generated a representation of the cluster which summarized the eye movement patterns of all the participants in one HMM. Figure 1 below shows the representation HMM and the fixation map of all the fixations combined. Figure 2 below shows the fixation maps per each fixation.

The left image in figure 1 shows the HMM model. For instance, the prior value of the red region suggests the probability of a first fixation belonging to that region. The probability of the next fixation transits from the red into the green region is 0.07.



	to red	to green	to blue
prior values	0.39	0.16	0.45
from red	0.69	0.07	0.24
from green	0.17	0.68	0.15
from blue	0.21	0.06	0.73

Figure 1: The image on the left shows the three GMM components of the HMM. Each colored region represents a ROI (red, green, or blue). The transition probabilities and the prior values are summarized in the table beneath. The image on the right shows the fixation map of all the fixations.

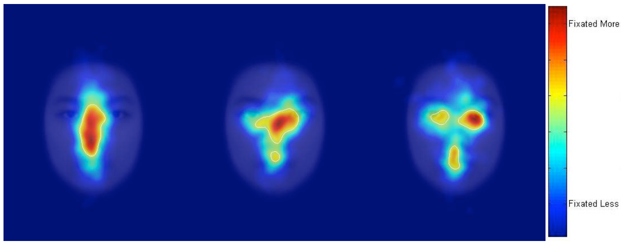


Figure 2: From the left to the right, the three images show the first, second, and third fixations that all subjects made.

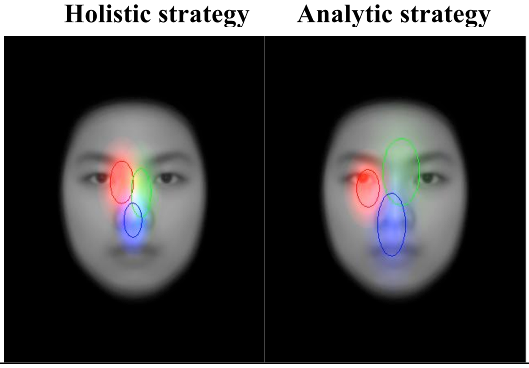
From the comparison between the VHEM output and the fixation map of all the fixations combined, it can be seen that the VHEM output was spatially similar to the fixation map. The fixation map showed that most fixations landed in the middle of the face, with some slightly to the right. The three Gaussian components found using the VHEM demonstrated a similar tendency. One advantage that the VHEM output has over the fixation map is that on top of the spatial distributions, it provides the temporal information of the eye movement data in the forms of the prior values and the transition probabilities.

The transition probabilities and the prior values suggested that in general, fixations were more likely to start from the red and the blue regions and to remain in or shift between the two regions. The chance of beginning from the green region was lower. However, these fixations were more likely to stay in the same region than moving to the other regions. The fixation maps are shown in Figure 2. While there appears to be some movement between fixations, the fixation maps carry no information about the actual saccade pattern. However, using the results from the HMM analysis, we can better interpret the fixation maps. The higher probabilities of remaining in the same regions and the lower probability of starting from the green region may have resulted in the fixations forming three separate clusters at the third fixation; the cluster corresponded to the green region was less compacted.

Section 1.2 - Two general strategies

Another advantage of using the HMM-based method is that the VHEM can group the input HMMs into several subgroups and generate a representation HMM for each subgroup. These would reveal the eye movement patterns shared by the participants in the same subgroup. The VHEM adopts a bottom-up, data-driven approach. It estimates the distance between an input HMM and a representation HMM. The distance between an input HMM and all the representation HMMs are then normalized, which gives a probability-based measure of how likely the input HMM belongs to a subgroup.

Using the VHEM, we discovered two subgroups, as shown in Figure 3 below.



Holistic	to red	to green	to blue
prior values	0.33	0.39	0.28
from red	0.65	0.24	0.11
from green	0.22	0.61	0.17
from blue	0.12	0.25	0.63
Analytic	to red	to green	to blue
prior values	0.06	0.47	0.47
from red	0.39	0.22	0.39
from green	0.03	0.75	0.22
from blue	0.05	0.25	0.70

Figure 3: The two representation HMMs of the two subgroups are shown in the left and the right images respectively.

It can be seen that the representation HMM on the left was more 'condensed'. The three Gaussian components were relatively small in size and were squeezed toward the center of the face. This pattern was similar to the "Eastern pattern" found in a previous study (Kelly et al., 2011) that was argued to represent a more holistic strategy. The HMM representation on the right, on the other hand, was more 'spread'. The three Gaussian components were large and more separated from one another. This pattern could be loosely associated with the "Western pattern" (Kelly et al., 2011) that represented a more analytic way of perceiving a face.

The table below shows the probabilities of the 32 HMMs belonging to the two subgroups. Each HMM was a model of a participant's eye movement patterns, so that the two numbers of each participant can be conceptualized as the degree to which the participant was biased to holistic or analytic eye movement strategies. Overall, 10 participants used holistic pattern, while 22 used the analytic strategy.

Table 1: Summary of the normalized log-likelihoods of the subjects belonging to the two subgroups.

ID	Holistic	Analytic	ID	Holistic	Analytic
01	0	1	17	0	1
02	0	1	18	.04	.96
03	1	0	19	0	1
04	1	0	20	1	0
05	0	1	21	1	0
06	0	1	22	0	1
07	1	0	23	0	1
08	0	1	24	0	1
09	0	1	25	0	1
10	1	0	26	1	0
11	0	1	27	0	1
12	0	1	28	1	0
13	1	0	29	1	0
14	0	1	30	.02	.98
15	0	1	31	0	1
16	0	1	32	0	1

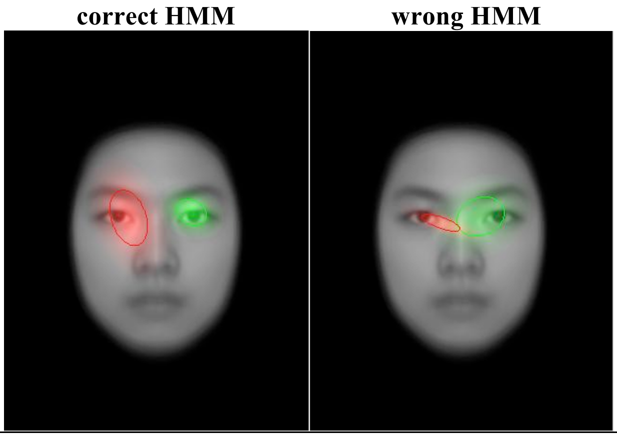
The log-likelihoods suggested that the two subgroups were very distinctive from each other. To confirm whether they really represented two distinctive eye movement patterns, we randomly created 50 pseudo-data chains; each was a sequence of three pseudo fixations. We measured the log-likelihoods of the two HMMs generating the pseudo-data. Paired t-test showed that the log-likelihoods generated by the two HMMs were significantly different, t (49) = -12.81, $p < .001$; mean log-likelihood difference was 13.84. The finding further confirmed that the two eye movement patterns were distinctive from each other.

Section 2 – Association between performance and eye movement patterns

To investigate whether the differences in recognition performance are associated with different eye movement patterns, we trained per participant an HMM on all the fixations collected from the correctly responded trials (correct HMM), and an HMM on all the fixations collected from the incorrectly responded trials (wrong HMM). We compared the mean log-likelihoods of the data being generated by the two HMMs.

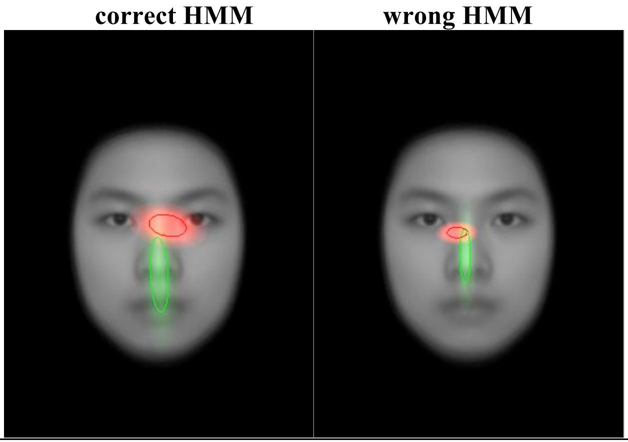
Paired t-test showed that the mean log-likelihoods of correct data being generated by the correct HMMs (M = -18.13) were significantly higher than the mean log-likelihoods of correct data being generated by the wrong

HMMs (M = -18.42), t (31) = -2.58, p = .01. The mean log-likelihoods of the wrong data being generated by the wrong HMMs (M = -17.9) was also significantly higher than the mean log-likelihoods of correct data being generated by the wrong HMMs (M = -18.53), t (31) = -4.58, p < .001. The results suggested that the two sets of HMMs were significantly different from each other. Figure 4 – 7 below illustrate the HMMs and the fixation maps of a few subjects.


correct HMM **wrong HMM**

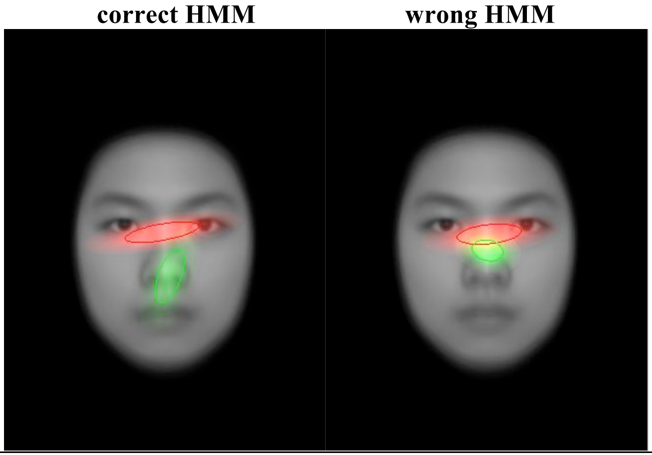
correct HMM	to red	to green
prior values	0.81	0.19
from red	0.19	0.81
from green	0.81	0.19
wrong HMM	to red	to green
prior values	0.50	0.50
from red	0.60	0.40
from green	0.87	0.13

Figure 4: The correct and wrong HMMs of subject 1.


correct HMM **wrong HMM**

correct HMM	to red	to green
prior values	0.61	0.39
from red	0.36	0.64
from green	0.66	0.34
wrong HMM	to red	to green
prior values	0.51	0.49
from red	0.54	0.46
from green	0.47	0.53

Figure 5: The correct and wrong HMMs for subject 2.


correct HMM **wrong HMM**

correct HMM	to red	to green
prior values	0.63	0.37
from red	0.68	0.32
from green	0.63	0.37
wrong HMM	to red	to green
prior values	0.39	0.61
from red	0.89	0.11
from green	0.83	0.17

Figure 6: The correct and wrong HMMs of subject 3.

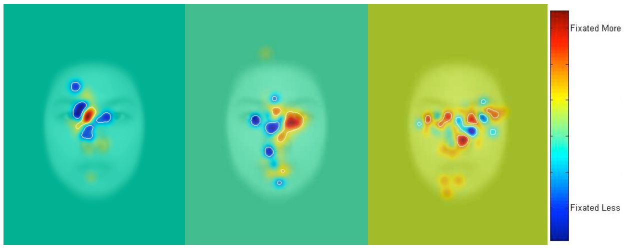


Figure 7: From the left to the right, the three images show the difference between the fixation maps of correct and the wrong responses of the three subjects shown in Figure 4-6.

From the figures above, we see that in some cases, the key difference between the wrong and correct HMMs can be discovered from the temporal rather than the spatial domain of the data. For instance, for subject 1, the correct and the wrong HMMs were spatially similar, but the wrong HMM had a different set of prior values and transition probabilities. If the subject started looking at the image from the right eye, the response is more likely to be incorrect. [1]

One disadvantage of comparing fixation maps between correct and wrong responses can be seen from figure 7 above. The pixel test in each case discovered many significantly different regions. These regions are all over the face, which make them very hard to be qualitatively explained.

Discussion

[1] We decided to restrict the correct and wrong HMMs to two hidden states because there was not enough data to train three hidden states in the wrong HMMs.

In this paper, we have proposed an HMM-based method to analyze eye movement data and demonstrated several advantages.

Firstly, our method can learn the ROIs for each person from the data together with their temporal information. This provides the information for describing and inferring the scan paths. Although fixation maps can be generated by fixations, such that the maps could be used to show the distributional difference of fixations over time, they do not contain transition information so that describing and inferring scan paths are impossible.

Secondly, using VHEM, the HMMs can be grouped into clusters based on their similarities. Our finding of this clustering showed that participants demonstrated either a holistic strategy or an analytic strategy. The two strategies were significantly different from each other.

Lastly, by comparing the correct and the wrong HMMs, we showed that the 'correct' eye movements were significantly different from the 'wrong' eye movements, and that the difference to a considerable extent can be attributed to the transition differences instead of spatial distribution differences. Comparison of the fixation maps of correct and wrong responses also showed the differences between the 'correct' and 'wrong' eye movements, but the differences were too spread so that the results lacked identifiable patterns. Also, the fixation map method was not able to show the difference in transition probability between eye movements in correct and wrong trials.

The lack of empirical findings to support the scan path theory caused eye movement researchers' lack of interest in sequential information (Henderson, 2003). Our findings, however, suggest that sequential information could be associated with performance. Theoretically, given a chain of fixations, using the two HMMs, the accuracy of the response can be predicted. This further justifies using HMMs to describe and analyze eye movement patterns. Future work will test this hypothesis.

In the current study, we pooled all the fixations together to find the ROIs because we assumed that the ROIs are the same across fixations. An alternative approach that does not rely on this assumption is to train the GMMs by fixation, so that at each fixation, there are a unique set of ROIs. An HMM in this case will have time-dependent states. For future research, we attempt to investigate this further.

In summary, here we show that eye movements can be better studied and understood using HMMs. With HMMs, we can describe both the spatial and the sequential aspects of eye movements. We also show that clustering the HMMs can yield interesting between-group differences. The two subgroups roughly correspond to more holistic and more analytic strategies. We further show that correct and wrong recognitions have different eye movement patterns and that the differences can be found in the transition probabilities.

Acknowledgements

A.B. Chan acknowledges support from the Research Grants Council of the Hong Kong SAR, China (CityU 110610). We thank Tina Liu for collecting the data for the study. We are also grateful to the Research Grant Council of Hong Kong (project code HKU 745210H to J.H. Hsiao) and the HKU Seed Funding Program for Basic Research (project code 201011159124 to J.H. Hsiao).

References

Baber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge: Cambridge University Press.

Barton, J. J. S., Radcliffe, N., Cherkasova, M. V., Edelman, J., & Intriligator, J. M. (2006). Information processing during face recognition: the effects of familiarity, inversion, and morphing on scanning fixations. *Perception 35*, 1089-1105.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Caldara, R. & Miellet, S. (2011). *iMap*: a novel method for statistical fixation mapping of eye movement data. *Behavior research methods 43*, 864-878.

Coviello, E., Chan, A. B., & Lanckriet, G. R. G. (2012). The variational hierarchical EM algorithm for clustering hidden Markov models. In *Neural Information Processing Systems* (NIPS).

Goldberg, J. H. & Helfman, J. I. (2010). Scanpath clustering and aggregation. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 227-234.

Gosselin, F. & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research 41*, 2261-2271.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *TRENDS in Cognitive Sciences 7*, 498-504.

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & Cognition 33*, 98-106.

Hsiao, J. & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological Science 19*, 998-1006.

Kelly, D. J., Jack, R. E., Miellet, S., De Luca, E., Foreman, K., & Caldara, R. (2011). Social experience does not abolish cultural diversity in eye movements. *Frontiers in Psychology 2*, 1-11.

Land, M. F. (2011). Oculomotor behavior in vertebrates and invertebrates. In S. Liversedge, I. Gilchrist, & S. Everling (Eds.), *Oxford handbook of eye movements*. Oxford: Oxford University Press.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation patterns made during brief examination of two dimensional images. *Perception, 26*, 1059-1072.

Pelphrey, K. A., Sasson, N. J., Reznick, S., Paul, G., Goldman, B. D., & Piven, J. (2002). Visual scanning of faces in autism. *Journal of Autism and Developmental Disorders 32*, 249-261.

Wade, N. J., & Tatler, B. W. (2011). Origins and applications of the eye movement research. In S. Liversedge, I. Gilchrist, & S. Everling (Eds.), *Oxford handbook of eye movements.* Oxford: Oxford University Press.