Understanding eye movements in face recognition using hidden Markov models

Tim Chuk Department of Psychology, University of Hong Kong, Hong Kong

Antoni B. Chan* Department of Computer Science, City University of Hong Kong, Hong Kong

Janet H. Hsiao* Department of Psychology, University of Hong Kong, Hong Kong

We use a hidden Markov model (HMM) based approach to analyze eye movement data in face recognition. HMMs are statistical models that are specialized in handling time-series data. We conducted a face recognition task with Asian participants, and model each participant's eye movement pattern with an HMM, which summarized the participant's scan paths in face recognition with both regions of interest and the transition probabilities among them. By clustering these HMMs, we showed that participants' eye movements could be categorized into holistic or analytic patterns, demonstrating significant individual differences even within the same culture. Participants with the analytic pattern had longer response times, but did not differ significantly in recognition accuracy from those with the holistic pattern. We also found that correct and wrong recognitions were associated with distinctive eye movement patterns; the difference between the two patterns lies in the transitions rather than locations of the fixations alone.

Introduction

How humans recognize faces has been of interest to researchers in different fields for several decades because faces are special in the sense that they provide viewers with unique and almost irreplaceable information about the identities, intentions, and emotions of other human beings. Some early studies (e.g., Ellis, Shepherd, & Davies, 1979; Young, Hay, McWeeny, Flude, & Ellis, 1985) showed that certain areas on a face are more informative than the others. For example, it was found that when the internal features (eyes, nose, and mouth) of familiar faces were shown to participants, their response time was faster than when the external features (hair and chin) were shown (Young et al., 1985). A more direct way to explore where people look at when they perceive faces is to track their eye movements. It was found that the eyes looked more frequently at the area between the eyebrows and the mouth (Walker-Smith, Gale, & Findlay, 1977), which covered the internal features of a face. This is a robust finding that has been replicated in many eye movement studies. For example, Mertens, Siegmund, and Grusser (1993) found that the preferred fixation targets on a face were the eyes, the nose, and the mouth. Luria and Strauss (1978) found that most fixations landed around the eyes, the nose, and the mouth; the ears, cheeks, and other parts of a face were fixated only when photographic negatives were presented (see Rayner, 1998, for a review).

Recent research suggests that eye movements in face recognition have functional roles and reflect underlying cognitive processes. For example, Henderson, Williams, and Falk (2005) showed that when participants were restricted to view face images only at the center of the faces, their recognition performances were significantly lowered than when they were allowed to view the face images freely. Hsiao and Cottrell (2008) showed that two fixations suffice in face recognition, and the two fixations are just around the center of the nose, demonstrating the preferred viewing location in face recognition. The preferred viewing location phenomenon has also been observed in eye movements in other cognitive tasks. For example, Rayner (1979) discovered that in reading English texts, the preferred viewing location is between the beginning and the middle of a word. Henderson (1993) found that in object viewing, the preferred viewing location is around the center of the object. These findings suggest that fixation loca-

Citation: Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *Journal of Vision*, *14*(11):8, 1–14, http://www.journalofvision.org/content/14/11/8, doi:10.1167/14.11.8.

doi: 10.1167/14.11.8 Received March 11, 2014; published September 16, 2014 ISSN 1534-7362 © 2014 ARVO

tions in cognitive tasks have functional roles and are not chosen by random selection.

Current eye movement analysis methods rely primarily on averaged data, which do not capture individual differences. However, recent research has shown substantial individual differences in eye movement behavior in cognitive tasks that persist over time. For example, in a scene perception study, Castelhano and Henderson (2008) found a high level of within-individual consistency in fixation duration and saccade amplitude when participants were viewing different types of visual stimuli. Risko, Anderson, Lanthier, and Kingstone (2012) discovered that personality may play a role in explaining individual differences in eye movement; they found that people who scored high on a curiosity measure significantly looked at more regions of a scene than those who scored low. Individual differences in eye fixation behavior during face identification have also been discovered. For example, Peterson and Eckstein (2013) found that individuals had different preferred fixation positions when viewing human faces and these individual preferences persisted over time. Kelly et al. (2011) discovered that people in different races have different eye movement patterns: Caucasians seem to look more toward the facial features (such as the eyes and the mouth) whereas Asians prefer to look at the center of the face (i.e., the nose). Thus, appropriate methods for analyzing eye movements are required to reflect individual differences.

In addition, most analysis methods focus on spatial information of eye movements, such as percentage of fixations in a set of predefined regions of interest (ROIs; e.g., Henderson et al., 2005), and do not take temporal information of eye movements such as the order of the ROIs visited into account. The ROI approach suffered from the lack of an objective method to define ROIs. For instance, in face recognition, Barton, Radcliffe, Cherkasova, Edelman, and Intriligator (2006) defined the mouth region as an irregularly shaped ROI around the mouth, while Henderson et al. (2005) defined the mouth region as a rectangular ROI that includes part of the cheek next to the mouth. Another problem is that predefined ROIs do not reflect individual differences in ROI. For example, in Barton et al. (2006), the upper part of the nose was included in the ROI with the two eyes; while this ROI may correctly capture the eye movements of some participants who consider the region as a potential target for eye movements, it may mistakenly interpret participants who looked only at the nose as fixating the eyes.

To address the problem of using pre-defined ROIs, more recent studies attempted to discover ROIs directly from data. A common method was to generate statistical fixation maps. A fixation map can be created by plotting the fixations and smoothing the map using a Gaussian function. Two fixation maps can be compared using a by-pixel test, which discovers statistically significant differences in pixels (Caldara & Miellet, 2011). For example, using fixation maps, it was found that the center of the nose was the most frequently viewed areas in face recognition (Hsiao & Cottrell, 2008). Through comparing two fixation maps, it was found that Caucasian participants prefer to look at the eyes and the mouth when viewing a face, whereas Asian participants tend to look at the nose (Kelly et al., 2011). Another method for finding ROIs using fixation maps is to cluster the centroids of significantly fixated regions using a clustering algorithm such as k-means clustering, and each resulting cluster represents an ROI (Jack, Blais, Scheepers, Schyns, & Caldara, 2009).

However, in reality eye movements are a sequence of saccades and fixations. The eyes fixate at a location shortly, before a saccade brings them to the next location. In this sense, eye movements may be considered as time-series data that are collected over time. Noton and Stark (1971) argued that during pattern perception, the eyes fixate at some features sequentially and this sequence of fixations, which they referred to as a scan path, will be repeated when the pattern is viewed again in the future (i.e., the scan path theory). The experiments testing the scan path theory found mixed results. For example, in a face perception task, participants showed regular scan paths for exactly the same images only about 65% of the time (Walker-Smith et al., 1977). In an image perception study, it was found that for some images, the between-individual similarities were higher than within-individual similarities, suggesting that the nature of the images influence how likely the individual-specific scan paths will be replicated (Josephson & Holmes, 2002).

Nevertheless, these findings should not be interpreted as undermining the significance of scan paths in cognitive tasks. In Laeng and Teodorescu (2002), participants were given pictures to look at and were then asked to imagine the pictures in front of a whiteboard while keeping their eyes open. In one condition, participants were told to move their eyes freely; in the other condition, they were told to fixate at the center of the board. It was found that in the free-viewing condition, participants performed significantly better in recalling the details of the pictures and that their scan paths resembled those when viewing the real images. It was suggested that when people tried to remember some visual contents, the muscle movements that brought the eyes fixating at different locations were encoded simultaneously so that scan paths served as aids to recalling the visual information.

Ellis and Smith (1985) suggested that although the exact scan paths may not be replicated, the order of fixations could be a statistically dependent stochastic process. Indeed, many studies have shown that saccades can be influenced by both top-down expectations (e.g., the task being performed; Yarbus, 1965) and bottom-up inputs (e.g., areas with salient features; Mannan, Ruddock, & Wooding, 1997). These findings imply that the target location of a saccade, i.e., the next fixation, is influenced by the location of the current fixation—specifically, the next fixation location can be considered as a random variable that takes one of several values with probabilities that are conditioned on the current fixation. In this sense, eye movements may be considered as a Markov stochastic process (a process where predictions of the future state depend only on the current state), which could be better understood using time-series probabilistic models.

In the current literature, the methods used to explore temporal information of eye movements also have shortcomings. For example, the string-editing method (e.g., Goldberg & Helfman, 2010) considers a scan path, i.e., a sequence of visited ROIs in eye movements, as a string, and the Levenshtein distance between two strings (i.e., the minimum number of steps to transform one string into the other string) is used as the measure of the temporal difference between two scan paths. However, this method relies on predefined ROIs, and is not able to reflect at which fixation two scan paths differ. A heat map-based solution is to generate fixation maps by fixation and compare between conditions (Caldara & Miellet, 2011). For instance, in an experiment with two conditions, the first fixations in each condition are used to generate two fixation maps. A comparison between the two fixation maps will show whether the two conditions differ significantly in the first fixation. However, the problem associated with this method is that areas showing significant differences are likely to be scattered so that the pattern can be hard to interpret, in particular when considering only one individual's data. In addition, transition information between fixations is lost in such maps. An alternative is to pool some of the significantly fixated areas on a group-averaged heat map together to form a group-level ROI using a clustering algorithm. The group-level ROIs can then be used as labels to convert a chain of fixations to a string. Regular fixation patterns of the group then can be inferred from these strings using the minimal description length method (Jack et al., 2009). This approach thus is data driven and considers both the spatial and temporal information. However, this method does not take individual differences in eye movements into account; also, the method assumes that all ROIs are circular and the same size, and the number of ROIs must be set by the experimenter (see more details in the Discussion).

Thus, to address the issues regarding individual differences and temporal information in eye movement data analysis, 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 in cognitive tasks. Previously, there were studies that used hidden Markov models to explore the relationship between eye movement and underlying cognitive mechanisms. For example, Liechty, Pieters, and Wedel (2003) applied HMM to model the shift between global and local covert attention when people were viewing advertisements. Their model consisted of two hidden states that corresponded to local and global attentions respectively. The eye movement data was considered as the emissions of the hidden states and therefore was used to estimate the parameters of the model. Simola, Salojarvi, and Kojo (2008) used HMM to discover the underlying cognitive processes in an online information search experiment. Participants were instructed to find certain words or sentences; their eye movements were assumed to imply what cognitive processes were taking place and were recorded to train the HMMs. They found three hidden states, which were interpreted as scanning, reading, and decision making, respectively.

In contrast to these studies, here we do not attempt to infer the underlying cognitive mechanism. Instead, we directly use HMMs to model eye movements; the hidden states of our model are directly associated with ROIs of eye movements. We conduct an eye movement study of face recognition and show that HMM is a data-driven approach that can (a) summarize an individual's general eye movement strategy, including person-specific ROIs and saccade patterns (i.e., transitions between the ROIs), (b) reveal between-subject similarities and differences of eye movement patterns, and (c) discover the association between recognition performance and eye movement patterns.

Methods

Materials

A total of 40 (20 males) grayscale frontal-view Asian faces were used. The faces were with neutral expressions and were unfamiliar to the participants. They were resized to maintain an interpupil distance of 80 pixels, and cropped to a standard size of 320 × 420 pixels. Participants viewed the screen at a distance of 60 cm; each face subtended about 6° of visual angle horizontally and 8° vertically. The size of the images was similar to seeing real faces at about 100 cm in distance. They were aligned by both vertical and horizontal eye positions.

Participants

We recruited 32 Asian participants (16 males), age ranged from 18 to 23, from the University of Hong

Kong. Participants were given course credit or honorariums. All participants reported normal or corrected-to-normal vision; they were all right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971).

Apparatus

An EyeLink 1000 eye tracker was used to monitor participants' eye movements. The tracking mode was pupil and corneal reflection with a sample rate of 2000 Hz. The standard nine-point calibration procedure was carried out before the experiment started and whenever the drift correction error was larger than 1° of visual angle. Stimuli were displayed on a 22-in. CRT monitor with a resolution of 1024 × 768 pixels and 150 Hz frame rate. A chin rest was used during the experiment.

Procedure

The experiment consisted of a training phase and a testing phase. In the training phase, participants were instructed to study 20 faces. In the testing phase, they were required to recognize the learned faces among 20 new faces. To counterbalance the asymmetry of the faces, half of the participants were tested with mirrored images.

At the beginning of each trial, a solid dot was shown at the center of the screen. Participants were told to fixate at the dot for drift correction. The dot was then replaced by a cross that stayed for 500 ms or until participants accurately fixated on it. In the training phase, the faces were presented either above or below the center of the screen for 5 s. Every two trials were separated by a one second blank. In the testing phase, participants were required to make their judgments by pressing corresponding buttons on a response pad using both hands. The faces were presented either above or below the center of the screen at random and stayed there until response. The mapping of the buttons was counterbalanced across participants. Before the experiment began, participants were given a practice session, in which they recognized three learned faces among three new faces, to get familiarized with the design. They were provided with feedback when they incorrectly responded during the practice.

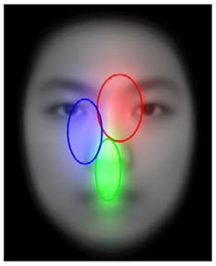
Hidden Markov model

Hidden Markov models (HMMs) are widely used to model data generated from Markov processes (Barber, 2012), i.e., a process whose next state is determined only by its current state. A state in an HMM is not directly observable; it can be inferred from the association between the assumed hidden state and the observed data, and the probability 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 representing 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. Finally, a vector of prior values indicates the probability of the process starting from a particular state (i.e., the initial state).

Here in the context of face recognition, each hidden state represents a different ROI of a face. The directly observable data is fixation locations, each belonging 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, i.e., the emission density for each hidden state is a two-dimensional Gaussian. The transition from the current hidden state to the next state represents a saccade, whose distribution is modeled by the transition matrix of the HMM. Finally, the ROI of the first fixation, i.e., the initial hidden state, is modeled by a categorical distribution (the prior values). We used a variational Bayesian approach to estimate the HMM parameters (Beal 2003, ch. 3; McGrory & Titterington, 2009), which placed prior distributions on the parameters of the emission density, transition matrix, and the initial state distribution, and factorized them to obtain the maximum a posteriori (MAP) estimate (Bishop, 2006). One advantage of this method was that it automatically determines the optimal number of hidden states (i.e., the number of ROIs), by pruning out unnecessary ones. Here we set the maximum number of hidden states to three and set the concentration parameter of the Dirichlet prior distributions on the initial state probabilities and transition matrix to 0.01. For the Normal-Wishart prior distribution on the Gaussian emissions, we set the prior mean to be the center of the image, and set the prior covariance matrix to be an isotropic covariance matrix with standard deviation of 14, which makes the prior ROI roughly the same size as the facial features.

We first trained one HMM per participant using fixations collected from all their trials to represent the participant's general eye movement pattern. To examine whether some participants had similar eye movement patterns, we used the variational hierarchical EM algorithm (VHEM) for HMMs (Coviello, Chan, & Lanckriet, 2012, 2014) to cluster the individual HMMs. For each cluster, the VHEM algorithm also produced a representation HMM (i.e., cluster center), which summarizes the common ROIs and transitions in the cluster.

To assess whether correct and wrong responses were associated with different scan patterns, we then trained two HMMs per participant using fixation sequences



Prior values	Red	Green	Blue
	0.47	0.35	0.17
Transition probabilities (from\to)	To Red	To Green	To Blue
From Red	0.60	0.21	0.19
From Green	0.42	0.56	0.02
From Blue	0.45	0.06	0.49

Figure 1. The representation HMM that summarized all the 32 individual HMMs using VHEM.

from all the trials with correct responses (i.e., correct HMM) and those with wrong responses (i.e., wrong HMM), respectively. To compare the correct and wrong HMMs, for each participant, we calculated the log-likelihoods of observing the fixation sequences that led to a correct response from the correct HMM, and then computed the mean. Similarly, we calculated the mean log-likelihood of observing the same sequences from the wrong HMM. Doing this on all the participants yielded two vectors of mean log-likelihoods of the correct and wrong HMMs respectively; we then examined whether the two vectors were significantly different. This procedure was repeated on the wrong fixation sequences. The difference between the two mean log-likelihoods was an approximation to the Kullback-Leibler (KL) divergence between the two corresponding HMMs, and was a measure of difference between two distributions.

Results

Behavioral data

The mean accuracy of the participants during the testing phase was 0.74 ($SD = 0.11$). Their mean reaction time was 1621.5 ms ($SD = 918.9$ ms). The average fixation duration was 261.78 ms ($SD = 46.11$ ms), and the average saccade length was 49.97 pixels ($SD = 13.56$ pixels).

Eye movement data

We used the VHEM algorithm to group the 32 HMMs (one for each participant) into one representation HMM that summarized all the individual HMMs. As shown in Figure 1, in general, a scan path most likely started from the red region (close to the right eye from the viewer's perspective), followed by the green region (close to the tip of the nose); it then either remained in or shifted between the two regions. The chance of beginning from the blue region (close to the left eye) was low, and once starting from the blue region (the left eye), the next fixation tended to either remain in the blue region or move to the red region (the right eye).

The VHEM output and the fixation map of all fixations (Figure 2) are spatially similar. The fixation map showed that the location slightly to the right of the bridge of the nose is the most fixated location (Hsiao & Cottrell, 2008)[1], which corresponded to the red region in our model. Nevertheless, although the by-fixation heat maps (Figure 2) show the changes of fixation distribution over time, they failed to capture the temporal dynamics of the scan paths. For instance, the by-fixation heat maps only show that most people started from the center of the face and explored the facial features later on; in contrast, in our model, the transition probability information among different ROIs show that about half of the participants who started from the lower center of a face (the green region) looked at the upper center of the face (the red region) next, whereas the other half remained looking at the lower center of the face (the green region). These seemingly two distinctive eye movement patterns can only be distinguished with the transition information among regions.

Two general strategies

We then clustered the individual HMMs into two subgroups using VHEM and generated a representation HMM for each subgroup. As shown in Figure 3, the top HMM is more condensed. Two regions are located in the center of the face; the other is slightly to the right. This pattern is similar to the holistic eye

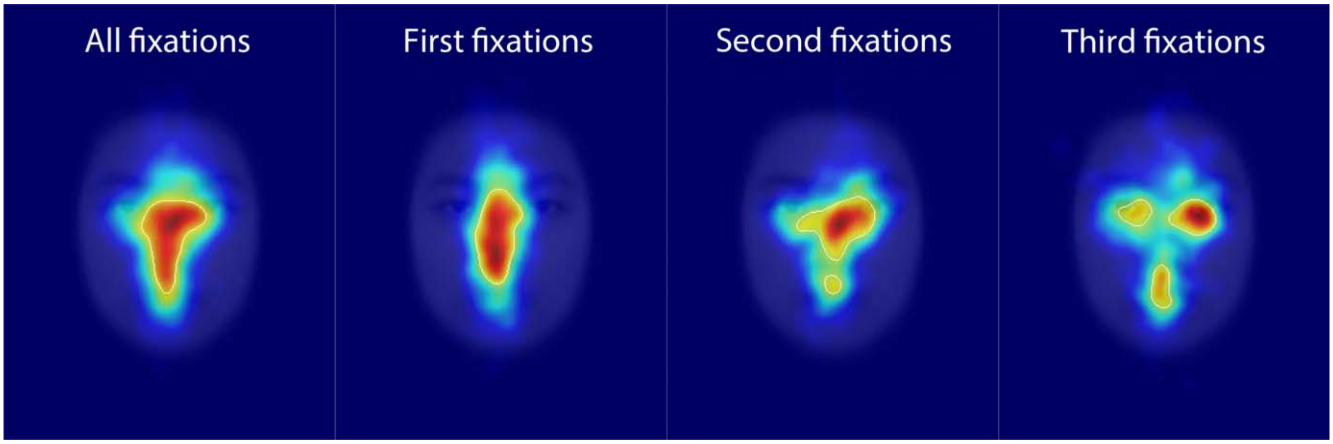
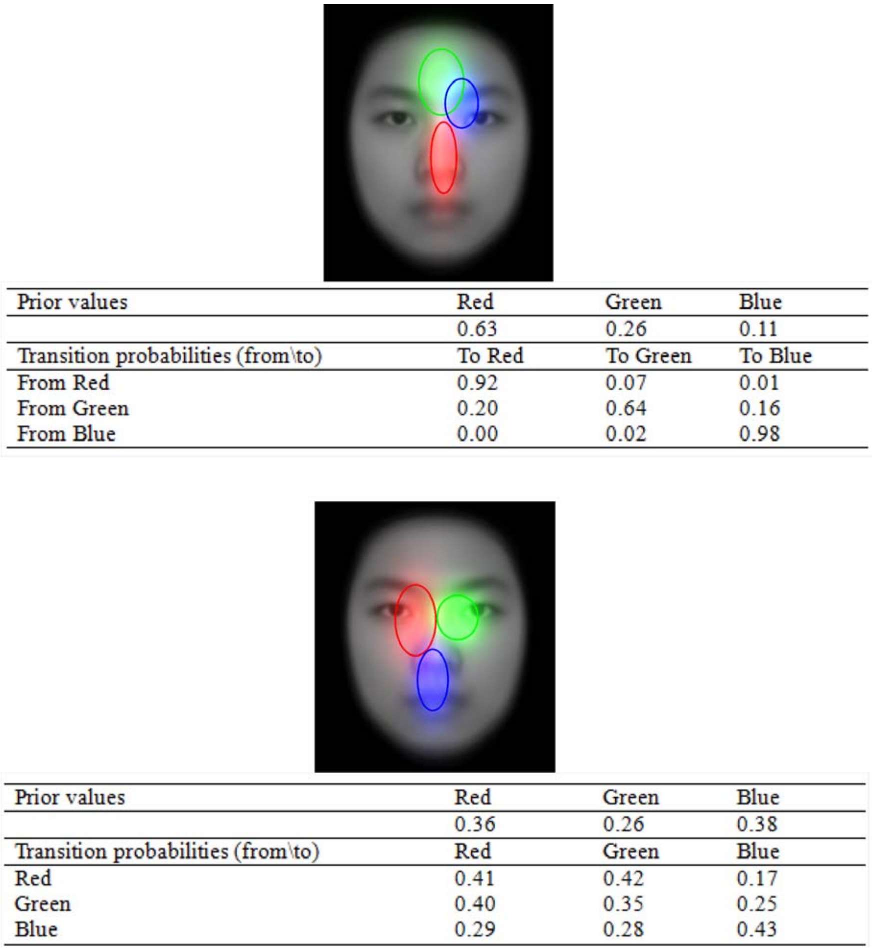


Figure 2. The heat maps of fixations.

movement pattern found in Asian participants in previous studies (Kelly et al., 2011). The transition matrix showed that participants in this subgroup preferred to start from the red region (i.e., the nose) and stay in the red region. In contrast, the HMM on the bottom is more spread out; the three regions are roughly on top of the left eye, the right eye, and the mouth respectively. This shows a more feature-based eye movement pattern similar to the analytic pattern observed in Caucasian participants (Kelly et al., 2011). The transition matrix showed that participants in this subgroup preferred to start from either the red or the



Prior values	Red	Green	Blue
	0.63	0.26	0.11
Transition probabilities (from\to)	To Red	To Green	To Blue
From Red	0.92	0.07	0.01
From Green	0.20	0.64	0.16
From Blue	0.00	0.02	0.98

Prior values	Red	Green	Blue
	0.36	0.26	0.38
Transition probabilities (from\to)	Red	Green	Blue
Red	0.41	0.42	0.17
Green	0.40	0.35	0.25
Blue	0.29	0.28	0.43

Figure 3. The two eye movement patterns. The one on top shows a holistic eye movement pattern whereas the one on the bottom shows an analytic eye movement pattern.

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

Table. The two columns indicate the probabilities of the individual HMMs belonging to the holistic and the analytic eye movement pattern groups.

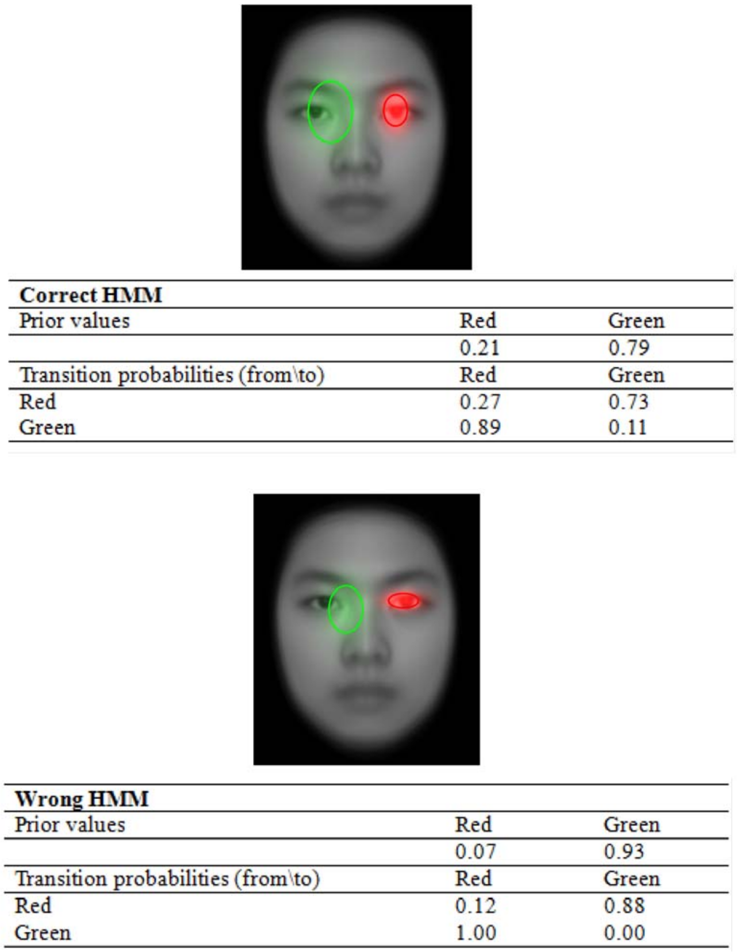
blue region and then move among the three regions. Since our participants were all Asians, this result shows that significant individual differences in eye movement patterns can be observed even within the same culture. Table 1 shows the probabilities of the 32 HMMs belonging to the holistic versus analytic subgroups. These probabilities could be conceptualized as the degree to which the participant was biased to holistic or analytic eye movement strategies. The results showed that about 65% of our participants adopted a relatively more analytic strategy.

To assess whether the holistic and analytic HMMs were significantly different from each other, we calculated the log-likelihoods of observing a holistic participant's fixation sequences from the holistic representation HMM and computed the mean. Similarly, we calculated the mean log-likelihood of observing the same sequences from the analytic representation HMM. Pairwise t test showed significant difference between the two vectors of mean log-likelihoods obtained from all participants, $t(10) = 3.44$, $p = 0.006$: Holistic participants' fixation sequences were significantly more likely to be generated by the holistic representation HMM than the analytic representation HMM. The same procedure was conducted on the analytic participants' sequences; pair-wise t test showed that the mean log-likelihoods generated by the two HMMs were also significantly different, $t(20) = -7.47$, $p < 0.001$: Analytic participants' fixation sequences were significantly more likely to be generated by the analytic than the holistic representation HMM. These results suggested that the holistic and analytic HMMs represented two distinctive eye movement patterns.

We also compared the recognition performance of the holistic and analytic groups. Analytic participants performed better (accuracy: 77%) than holistic participants (accuracy: 70%); nevertheless, this difference did not reach statistical significance, $t(30) = 1.64$, $p = 0.11$. The analytic participants had longer reaction times (1751.3 ms) than the holistic participants (1373.5 ms), $t(30) = 2.30$, $p = 0.03$. This finding stood when we only considered response times of correct trials (analytic participants: 1674.9 ms vs. holistic participants: 1303.3 ms), $t(30) = 2.57$, $p = 0.02$. Our further analysis showed that analytic participants produced more fixations than holistic participants, $t(30) = 2.11$, $p = 0.04$. In addition, analytic participants had longer average saccade lengths (mean = 54.81 pixels) than holistic participants (mean = 40.75 pixels), $t(30) = -3.17$, $p = 0.004$. Nevertheless, there was no significant difference in average fixation duration between the analytic (mean = 255.41 ms) and the holistic participants (mean = 273.94 ms), $t(30) = -1.08$, $p = 0.29$. These findings showed that participants who adopted an analytic/feature-based strategy on average made more fixations, made longer saccades, and spent more time when recognizing faces. Note that the clustering algorithm (VHEM) only used information available in the individual HMMs, that is, fixation locations and transitions probabilities, without any behavioral data; the difference in response time and number of fixations between the holistic and analytic groups emerged naturally, reflecting the fundamental differences between the two eye movement strategies.

Association between performance and eye movement patterns

We then investigated whether within each individual the correct and wrong responses were associated with different eye movement patterns. We trained per participant one HMM using all fixation sequences from the trials with correct responses (correct HMM), and



Correct HMM		
Prior values	Red	Green
	0.21	0.79
Transition probabilities (from\to)	Red	Green
Red	0.27	0.73
Green	0.89	0.11

Wrong HMM		
Prior values	Red	Green
	0.07	0.93
Transition probabilities (from\to)	Red	Green
Red	0.12	0.88
Green	1.00	0.00

Figure 4. The correct and the wrong HMMs of one individual.

another HMM using all sequences from those with wrong responses (wrong HMM). Using the method introduced in the Methods section, we found that the mean log-likelihoods of the sequences that led to correct responses being generated by the correct HMMs ($M = -17.68$) were significantly higher than the mean log-likelihoods of the same correct sequences being generated by the wrong HMMs ($M = -22.22$), $t(31) = 5.04$, $p < 0.001$. In other words, participants' fixation sequences that led to a correct response were significantly more likely to be generated by the correct HMM than the wrong HMM. Similarly, the mean log-likelihoods of the sequences that led to wrong responses being generated by the wrong HMMs ($M = -10.56$) was also significantly higher than the mean log-likelihoods of the same wrong sequences being generated by the correct HMMs ($M = -11.04$), $t(31) = -3.39$, $p = 0.002$. These results showed that the correct and wrong HMMs were significantly different, suggesting that fixation sequences that led to wrong and correct responses were distinctive from each other.

In addition, we discovered that in some participants, the major difference between the correct and wrong HMMs was in their transition information rather than spatial distributions of the ROIs. As illustrated in Figure 4, in this participant, the spatial distribution of the ROIs in the correct HMM resembled that of the ROIs in the wrong HMM. The prior values and the transition probabilities, however, were distinctive from each other. More specifically, this participant started from the green region less frequently in correct trials than in wrong trials and was more likely to switch between the red and green regions in correct trials than in wrong trials.

The heat map approach compares the correct and wrong fixation patterns by generating two heat maps, respectively, and performing a pixel-by-pixel comparison (Caldara & Miellet, 2011). Figure 5 shows the output of the comparison of the same individual in Figure 4. It can be seen that the areas showing significant difference (circled in white) are very scattered, so that the difference between the correct and wrong fixation patterns is hard to interpret.

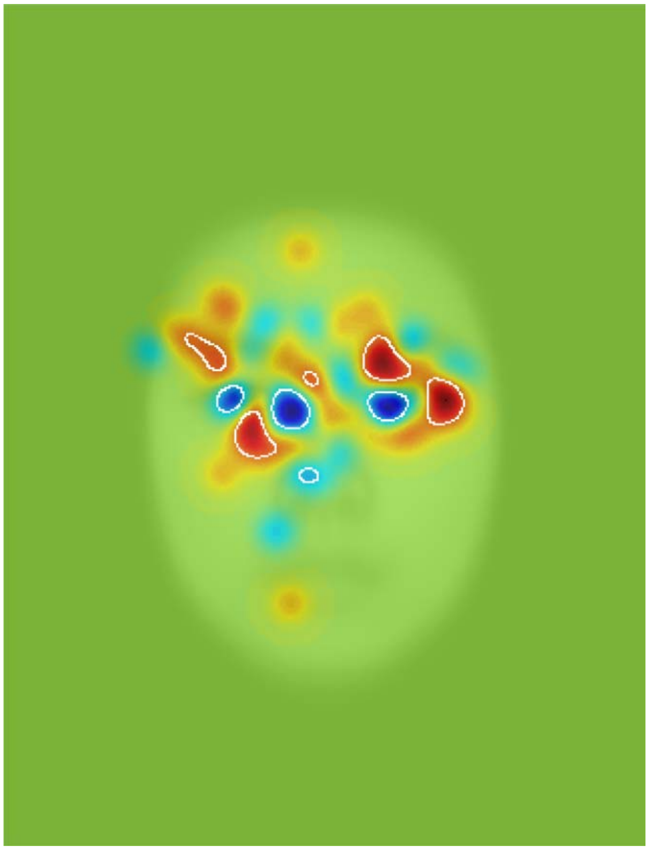


Figure 5. The difference-heat map of the same individual in Figure 4. The areas surrounded by white boundaries are those with significant difference between the two conditions.

To assess the importance of the transition information of the HMMs in accounting for the difference in eye movement pattern between the correct and wrong trials, we compared the ability of the HMM models with and without transition information in predicting whether a scan path led to a correct or wrong response. We removed the transition information of the HMMs by assuming that the fixations were statistically independent from each other, i.e., a fixation has no impact on the preceding and the following fixations; this was done by marginalizing out the previous states. Hence, an HMM was turned into a set of Gaussian mixture models (GMMs), each representing the distribution of one fixation, without transition probability information among regions. This analysis is analogous to the by-fixation method of analyzing temporal differences in the heat map approach (Caldara & Miellet, 2011). For each fixation sequence that led to a correct response, we calculated the log-likelihood of the sequence being generated by the correct GMMs as the sum of the log-likelihoods of observing each fixation from its corresponding GMM. Then we calculated the log-likelihoods of observing the same sequence using the wrong GMMs. For a sequence that led to a correct response, if the correct GMMs produced a higher log-likelihood than the wrong GMMs (which indicated that the fixation sequence was more likely to be generated by the correct GMMs than the wrong GMMs), we considered it as a correct prediction. Similarly, if the wrong GMMs produced a higher log-likelihood than the correct GMMs for fixation sequences that led to a wrong response, we considered it as a correct prediction. We found that the average accuracy of the marginalized models (GMMs, i.e., the models without transition information) in predicting the responses was 0.62; in contrast, the average accuracy of the same analysis using the HMMs (i.e., the models with transition information) was 0.73, significantly higher than that of the marginalized models (GMMs), $t(31) = 7.63$, $p < 0.001$. This finding further demonstrated that the difference in eye movement pattern between correct and wrong responses lies in the transitions rather than the locations of the fixations alone.

Discussion

Here we have proposed an HMM based approach for analyzing eye movement data. We applied the method to eye movements in face recognition and obtained several important findings that were not possible with previous analysis methods.

In contrast to the previous methods, our HMM approach has demonstrated several advantages. Firstly, our method can learn the ROIs for each person from the data together with their temporal information, i.e., the transitions from one ROI to another. This provides the information for describing and inferring the scan paths. Although fixation maps can be generated separately for each fixation to show the distributional difference of fixations over time, they do not contain transition information among different regions of the maps so that describing and inferring scan paths are impossible.

Secondly, using the clustering algorithm 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 (i.e., mainly looking at the center of a face) or an analytic strategy (i.e., a preference over looking at the eyes and mouth of a face) similar to those observed in Asians and Caucasians respectively in previous studies (Blais, Jack, Scheepers, Fiset, & Caldara, 2008; Kelly et al., 2011). The two strategies were significantly different from each other: Eye movements of the participants using the holistic strategy were significantly more likely to be generated by the holistic HMM, whereas those of the participants using the analytic strategy were significantly more likely to be generated by the analytic

HMM. This result suggests that even within the same culture, there are significant individual differences in eye movement strategies. About 60% of our Asian participants exhibited relatively more analytic eye movement patterns. Note that such clustering is completely driven by the eye movement data alone (in contrast to Miellet, Caldara, & Schyns, 2011, in which they grouped participants into either a global or local strategy according their behavior in identifying faces in which local/foveated and global/parafoveal information belonged to different famous identities). In addition, we showed that participants using an analytic eye movement strategy in general had longer response times and made more fixations in face recognition than those using a holistic strategy; nevertheless, the two groups did not differ significantly in recognition accuracy. As our clustering algorithm used only eye movement data (i.e., the locations of the ROIs and the transition probabilities among them) available in the HMMs, these differences in face recognition behavior between the two strategies emerged naturally as the result of such clustering.

Lastly, by comparing the correct and the wrong HMMs, we showed that eye movements that led to correct responses were significantly different from those that led to wrong responses; the difference to a considerable extent is attributed to the transition differences among fixation locations, instead of spatial distribution differences alone. Thus, taken together, our data suggest that individuals have different preferred viewing locations in face recognition (such as the holistic vs. analytic strategies), consistent with recent studies showing substantial individual differences in eye fixation behavior in cognitive tasks that persist over time (e.g., Castelhano & Henderson, 2008; Peterson & Eckstein, 2013). In addition, the order of visiting these preferred viewing locations can be different depending on the individual's mental state, i.e., whether he/she is likely to make a correct or wrong response. Indeed, we showed that given a chain of fixations, the participant's response accuracy could be predicted by comparing the likelihood of the chain of fixations being generated by the correct and the wrong HMMs; with only spatial information, the average accuracy of the models was significantly worse than that of the models using both spatial and temporal information. This finding showed that temporal information should not be neglected and demonstrated the advantage of analyzing temporal and spatial information as a whole in eye movement data analysis. It also further justifies using HMMs to analyze eye movement patterns. In contrast, although a comparison of the fixation maps of correct and wrong responses also showed the differences between the correct and wrong eye movements, the differences were so spread 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.

Thus, our HMM based approach provides an objective data-driven method to examine both between-subject and within-subject differences in eye movement pattern, using both spatial and temporal information of the eye movements. This is in contrast to the previous methods, in which typically the spatial and the temporal information are considered separately, and individual differences are considered noise. Also, our method does not rely on predefined information. Unlike previous studies that used pre-allocated ROIs, in our method the spatial information (ROIs) can be learned directly from data. This is particularly useful for tasks in which ROIs are unknown or cannot be accurately predefined due to individual differences. For instance, the ROIs on a webpage are typically ill defined and task dependent, (Eckhardt et al., 2013). Even when ROIs can be relatively clearly defined, such as in face recognition, using predefined ROIs is likely to introduce experimenter bias; in addition, it does not reflect individual differences in ROIs. When examining between-subject differences, conventional methods usually require a criterion to divide participants into groups (e.g., Kelly et al., 2011; Miellet et al., 2011). In contrast, groups with different eye movement patterns can be discovered directly from data using our approach.

In our HMM based approach, we use HMM to directly model the eye movement patterns of our participants. The hidden states are assumed to be the ROIs and the eye movement data is assumed to be emissions of the ROIs. In contrast to this approach, HMM can also be used to model cognitive states underlying eye movement data. For example, in Liechty et al. (2003), the two hidden states represented global and local covert attention. In Simola et al. (2008), the three hidden states represented scanning, reading, and decision making. Hayashi (2003) used HMMs to model pilots' instrument scanning behavior, and the hidden states of the HMMs corresponded to different instrument tracking tasks (e.g., vertical, horizontal, or airspeed-tracking tasks); it was discovered that unlike less experienced pilots, expert pilots monitor more instrument tracking tasks when flying under low visibility. Future work will examine the possibility of using HMM to model different cognitive states in face recognition, and the possibility of using hierarchical HMM to simultaneously model eye movement behavior and cognitive states.

In addition to HMM approaches, there are other analysis methods that can potentially take both temporal and spatial information of eye movements into account. For example, Jack et al. (2009) analyzed Caucasian and Asian participants' eye movement data in a facial expression judgment task. They analyzed the spatial information of the eye movements by applying a pixel test on a fixation map (heat map) to find the significantly fixated areas and their centroids. In order to establish a set of common ROIs across conditions for comparison, they pooled all the centroids from the fixation maps in different conditions and ran a k-means clustering to find a centroid for each nonoverlapping fixated region. This approach offers data-driven, clearly defined ROIs. Note however that the k-means clustering algorithm assumes that each cluster is isotropic and all clusters have the same spatial size (formally, each cluster is a Gaussian with the same isotropic covariance matrix), and hence the resulting ROIs will tend to be circular and have similar sizes. In contrast, the HMM approach we used here allows each ROI to have a different shape and size (i.e., each ROI is described by a Gaussian with its own covariance matrix). Thus, the HMM approach offers a more flexible representation of ROIs. In addition, the k-means approach requires the number of ROIs to be set explicitly by the experimenter, whereas our Bayesian HMM automatically selects the number of ROIs based on the observed data and the prior. Although the parameters for the Bayesian prior are still selected by the experimenter, the prior only indirectly affects the number of ROIs (in general, the Bayesian HMM will find the same number of ROIs over a range of prior parameters). As for the temporal information of the eye movements, after obtaining the ROIs, Jack et al. (2009) used the ROIs to describe sequences of fixations as strings, and then used the minimum description length method to extract regular patterns from the data. Note that in describing a fixation sequence as a string, each fixation is assigned to only one ROI (i.e., a "hard" assignment). However, a "hard" assignment may not be suitable for a fixation that is close to the boundary between two ROIs, since a small amount of noise (e.g., due to measurement error in the eye tracker) could have caused the fixation to be categorized differently, resulting in a different string and perhaps confounding the subsequent analysis. In contrast, when estimating the ROIs and transition matrices, our HMM uses "soft" assignments, where each fixation is associated with a posterior probability of belonging to each ROI. Hence, a fixation on a ROI boundary is considered to belong equally to both ROIs, thus reducing potential problems due to these ambiguous fixations. When extracting regular patterns, Jack et al. (2009) considered fixation patterns from zeroth (single fixation sequences) up to third order (four fixation sequences). In contrast, the transition matrix of the HMM used here only describes the first-order transition probabilities between any two ROIs and does not contain information about the whole sequence. Nevertheless, the HMM approach can be expanded to incorporate higher order transition probabilities (i.e., higher order HMMs). Finally, in Jack et al. (2009), the analysis of spatial and temporal information of eye movements were conducted as two separate parts, which is suboptimal from an estimation standpoint, whereas in our HMM approach the two types of information are analyzed simultaneously. In addition, our HMM approach is able to take individual differences into account in group/condition comparisons.

Recent research has suggested that in describing one's eye fixation patterns, fixation duration information or saccade length information can be important in some cases. For instance, it was found that autistic patients viewed the nonfeature parts of a face significantly longer and the facial features significantly shorter than normal people (Pelphrey et al., 2002). In the current study, we only used fixation location information in the HMMs. Future work will incorporate the duration information and saccade length information as additional dimensions of the hidden state emissions in the HMM to more accurately model participants' eye fixation behavior.

Although in the current study, we observed two groups of participants with different eye fixation patterns (holistic vs. analytic patterns), whether the difference in eye fixation pattern leads to different uses of information remains unclear. More specifically, although eye movement data help us understand how faces are perceived, there is no direct association between where the eyes land and what information are extracted and processed. In many studies, it was found that the most frequently viewed area is the center of the nose. However, this does not mean that only the center of the nose was viewed and processed; participants might obtain a peripheral view of the other parts of the face, or engage global attention to obtain a holistic representation of the face. The reverse-correlation methods, such as the Bubbles (Gosselin & Schyns, 2001), have been proposed to discover diagnostic features used by participants for a recognition task. More specifically, in the Bubbles technique, bubbles (Gaussian-shaped windows) of different spatial frequency and sizes are used to reveal only parts of a face in a face recognition task; through examining the characteristics of the bubbles that lead to a correct response, it is possible to infer diagnostic features that are used by participants. To understand the link between eye fixation behavior and information use in face recognition, future work will examine whether people with different eye movement patterns, as identified by our HMM based method, use different diagnostic features for face recognition through the reverse-correlation methods (see Jack et al., 2009).

In summary, here we show that eye movements can be better understood using HMMs. With HMMs, we

can describe both the spatial and sequential aspects of eye movements; in addition, clustering the HMMs can yield interesting between-group differences. Through applying this method to eye movements in face recognition, we show that participants' eye movement patterns can be clustered into two subgroups that roughly correspond to holistic and analytic strategies, and participants using an analytic strategy generally make more fixations and have longer response times in face recognition. We also show that within an individual correct and wrong recognitions have different eye movement patterns, and that the difference lies in the transitions rather than the locations of the fixations alone. These findings are not possible with previous methods that do not take individual differences and temporal information into account.

Keywords: hidden Markov models, face recognition, eye movement

Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (project code HKU 745210H to J. H. Hsiao, and project code CityU 110513 to A. B. Chan). We thank the editor, Prof. Keith Rayner, and two anonymous reviewers for helpful comments. We also thank Emanuele Coviello for providing the VHEM code and Tina Liu for collecting the behavioral data.

*AC and JH contributed equally to this article.
Commercial relationships: none.
Corresponding author: Tim Chuk.
Email: u3002534@connect.hku.hk.
Address: Department of Psychology, University of Hong Kong, Hong Kong.

Footnote

[1] In contrast, with Caucasian participants Hsiao and Cottrell (2008) showed that the most fixated region was slightly to the left of the bridge of the nose (see also Butler et al., 2005). Whether this difference is related to culture differences requires further examinations (see also Saether, Van Belle, Laeng, Brennen, & Øvervoll, 2009).

References

Barber, 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.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference* (Doctoral dissertation). University of London.

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

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PLoS ONE, 3*, e3022. doi:10.1371/hournal.pone.0003022.

Butler, S., Gilchrist, I. D., Burt, D. M., Perrett, D. I., Jones, E., & Harvey, M. (2005). Are the perceptual biases found in chimeric face processing reflected in eye-movement patterns? *Neuropsychologia, 43,* 52–59.

Caldara, R., & Miellet, S. (2011). iMap: A novel method for statistical fixation mapping of eye movement data. *Behavior Research Methods, 43,* 864–878.

Castelhano, M. S., & Henderson, J. M. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology, 62,* 1–14.

Coviello, E., Lanckriet, G. R., & Chan, A. B. (2012). The variational hierarchical EM algorithm for clustering hidden Markov models. In P. Barlett (Eds.), *Advances in neural information processing systems* (pp. 404–412). New York: Curran Associates, Inc.

Coviello, E., Chan, A. B., & Lanckriet, G. R. G. (2014). Clustering hidden Markov models with variational HEM. *Journal of Machine Learning Research, 15*(Feb), 697–747.

Eckhardt, A., Maier, C., Hsieh, J. J., Chuk, T., Chan, T., Hsiao, J., & Buettner, R. (2013). Objective measures of IS usage behavior under conditions of experience and pressure using eye fixation data. *Proceedings of 34th International Conference on Information Systems (ICIS)*, December 16–18, 2013, Milan, Italy.

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception, 8,* 431–439.

Ellis, S. R., & Smith, J. D. (1985). Patterns of statistical dependency in visual scanning. In R. Groner, G. W. Mcconkie & C. Menz, (Eds)., *Eye movements and human information processing* (pp. 221–238). Amsterdam: Elsevier Science Publishers BV.

Goldberg, J. H., & Helfman, J. I. (2010). Scanpath clustering and aggregation. *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 227–234), March 22–23, 2010, Austin, TX, USA.

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

Hayashi, M. (2003). Hidden Markov models to identify pilot instrument scanning and attention patterns. *Systems, Man and Cybernetics, 3,* 2889–2896.

Henderson, J. M. (1993). Eye movement control during visual object processing: Effects of initial fixation position and semantic constraint. *Canadian Journal of Experimental Psychology, 47,* 79–98.

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.

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Current Biology, 19,* 1–6.

Josephson, S., & Holmes, M. E. (2002). Attention to repeated images on the World-Wide Web: Another look at scanpath theory. *Behavior Research Methods, Instruments, & Computers, 34,* 539–548.

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.

Laeng, B., & Teodorescu, D.-S. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science, 26,* 207–231.

Liechty, J., Pieters, R., & Wedel, M. (2003). Global and local covert visual attention: Evidence from a Bayesian hidden Markov model. *Psychometrika, 68,* 519–541.

Luria, S. M., & Strauss, M.S. (1978). Comparison of eye movements over faces in photographic positives and negatives. *Perception, 7,* 349–358.

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

McGrory, C. A., & Titterington, D. M. (2009). Variational Bayesian analysis for hidden Markov models. *Australian & New Zealand Journal of Statistics*, *51*(2), 227–244.

Mertens, I., Siegmund, H., & Grusser, O. J. (1993). Gaze motor asymmetries in the perception of faces during a memory task. *Neuropsychologia, 31,* 989–998.

Miellet, S., Caldara, R., & Schyns, P. G. (2011). Local Jekyll and global Hyde: The dual identity of face identification. *Psychological Science, 22,* 1518–1526.

Noton, D., & Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research, 11,* 929–942.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9,* 97–113.

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

Peterson, M. F., & Eckstein, M. P. (2013). Individual differences in eye movements during face identification reflect observer specific optimal points of fixation. *Psychological Science*, *24*(7), 1216–1225.

Pieters, R., Rosbergen, E., & Wedel, M. (1999). Visual attention to repeated print advertising: A test of scanpath theory. *Journal of Marketing Research, 36,* 424–438.

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception, 8,* 21–30.

Risko, E. F., Anderson, N. C., Lanthier, S., & Kingstone, A. (2012). Curious eyes: Individual differences in personality predict eye movement behavior in scene-viewing. *Cognition, 122,* 86–90.

Saether, L., Van Belle, W., Laeng, B., Brennen, T., & Øvervoll, M. (2009). Anchoring gaze when categorizing faces' sex: Evidence from eye-tracking data. *Vision Research, 49,* 2870–2880.

Simola, J., Salojarvi, J., & Kojo, I. (2008). Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research, 9,* 237–251.

Walker-Smith, G. J., Gale, A. G., & Findlay, J. M.

(1977). Eye movement strategies involved in face perception. *Perception, 6,* 313–326.

Yarbus, A. L. (1965). *Eye movements and vision, translated from Russian by Basil Haigh*. New York: Plenum Press.

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception, 14,* 737–746.