The Role of Eye Movement Consistency in Learning to Recognise Faces: Computational and Experimental Examinations

Janet H. Hsiao (jhsiao@hku.hk)
Department of Psychology and the State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong
Pokfulam Road, Hong Kong

Jeehye An (anjeehye@connect.hku.hk)
Department of Psychology, University of Hong Kong
Pokfulam Road, Hong Kong

Antoni B. Chan (abchan@cs.cityu.edu.hk)
Department of Computer Science, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong

Abstract

In face recognition, the frequency of looking at the eyes, the most diagnostic feature, predicts better performance in adults but not in children, suggesting that different factors may underlie children's face recognition performance. Here we test the hypothesis that eye movement consistency plays an important role during early learning stages. Through computational modelling that combines a deep neural network and a hidden Markov model that learns eye movement strategies by interacting with the network, we showed that consistency instead of eye movement pattern better predicted face recognition performance during early learning stages. Similarly, in human studies, children's consistency but not pattern of eye movements predicted face recognition performance, and their eye movement consistency was associated with executive function abilities. Thus, learning to recognise faces initially involves developing a consistent visual routine, which depends on executive function abilities. This finding has important implications for learning in both healthy and clinical populations.

Keywords: Eye movement; face recognition; deep neural network; hidden Markov model; entropy

Introduction

In face recognition, the eyes are the most diagnostic features, and adults who look at the eyes more often have better performance (Chuk, Crookes, Hayward, Chan, & Hsiao, 2017; Vinette, Gosselin, & Schyns, 2004). Nevertheless, in children the frequency of looking at the eyes does not predict better performance; also, children with Autism Spectrum Disorders (ASDs) did not differ from matched controls in the frequency of looking at the eyes in static face recognition regardless of their poorer performance (Wilson, Palermo, & Brock, 2012). Thus, factors other than looking at the eyes/diagnostic information may play a more important role during early stages of learning.

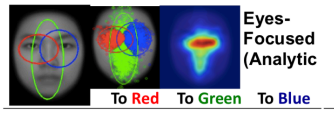
Adults are shown to have observer-specific fixation behaviour in face recognition that persists over time, and deviation from this visual routine results in suboptimal performance (Peterson & Eckstein, 2013). This phenomenon may be because visual routines facilitate extraction of learned diagnostic features (perceptual learning; Nazir & O'Regan, 1990). Inconsistent eye movements during early learning may reflect difficulty in discovering and extracting diagnostic features to develop a visual routine, resulting in suboptimal performance. The discovery and extraction of diagnostic features may depend on cognitive abilities. Thus, early learners/children may have less consistent eye movements than experts/adults, and eye movement consistency may better predict their performance. Here we tested this hypothesis through both computational and experimental examinations. Computational modelling enables manipulation of factors that are difficult to control in human subjects, such as maturation difference between children and adults. It also offers explanations and predictions for human behaviour. We then conducted a human study to examine the predictions.

The advance of deep neural networks (DNNs) has revolutionized the research on automatic face recognition (Wang & Deng, 2019) and cognitive modelling (Yamins et al., 2014). Nevertheless, DNNs typically assume that all aspects of the input can be processed simultaneously for efficiency and accuracy. This differs from how humans recognize visual objects through a sequence of eye fixations. Previous models of human face recognition that take eye fixations into account typically only use bottom-up salience-based measures for fixation selection, and are not designed to model eye movement strategy learning (Barrington, Marks, Hsiao & Cottrell, 2008). More recent models simulate top-down visual attention by using the internal representation of a DNN at previous time steps to predict the next attended location/object for image captioning (Ablavatski, Lu, & Cai, 2017). However, their attention mechanism has been developed mainly for object detection in a cluttered scene. Similarly, Mnih, Heess, Graves, and Kavukcuoglu (2014) proposed a recurrent network for visual attention for finding digits in a cluttered image. To our knowledge, no previous model has implemented a top-down attention mechanism for

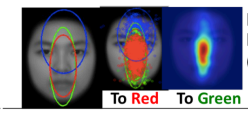
©2020 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

learning eye movement strategies for visual object/face recognition through integrating information across multiple fixations using deep learning.

Here we proposed a novel computational model that combines a DNN and a hidden Markov model (HMM) to learn eye movement strategies including both sequences of fixation locations and associated attention window sizes (global/local attention) for recognition. The DNN learns optimal perceptual representations under the guidance of an attention mechanism summarized in an HMM, and the HMM learns optimal eye movement strategies through feedback from the DNN. In contrast to previous approaches that have an arbitrary attention mechanism, here we assume that fixations occur within subject-specific ROIs, since we are interested in individual differences in which object/facial features are important for recognition. We also include constraints of human perception, such as saccade noise and visual-spatial acuity of the retina, into our computational model. HMM is a statistical time-series model commonly used to model eye movement data. In particular, the Eye Movement analysis with Hidden Markov Models (EMHMM) method has recently been proposed for summarizing and quantifying an individual's eye movement pattern (Chuk, Chan, & Hsiao, 2014). Specifically, a person's eye movements can be modelled in terms of both person-specific regions of interest (ROIs) and transitions among the ROIs using an HMM. The hidden states correspond to the ROIs; parameters are estimated directly from data using a variational Bayesian approach that can automatically determine the optimal number of ROIs of the model. Individual HMMs can be clustered using the variational hierarchical EM algorithm (Coviello, Chan, & Lanckriet, 2014) to reveal common patterns, such as the eyes-focused and nose-focused patterns in face recognition (Chan, Chan, Lee, & Hsiao, 2018; Figure 1). Similarities among individual patterns can be assessed quantitatively using data likelihood measures. Thus, this method is particularly suitable for examining the relationship between eye movements and other measures.



	To Red	To Green	To Blue
Priors	.03	.82	.15
From Red	.21	.34	.45
From Green	.25	.36	.39
From Blue	.27	.26	.47



	To Red	To Green	To Blue
Priors	.16	.80	.04
From Red	.82	.12	.06
From Green	.93	.02	.05
From Blue	.07	.00	.93

Figure 1: Eyes-focused and nose-focused eye movement patterns discovered in Chan et al. (2018).

EMHMM has been applied to face recognition research and uncovered novel findings not revealed by other methods. For example, the eyes-focused pattern was associated with better recognition performance than the nose-focused pattern (Chan et al., 2018). Also, individuals have preferred eye movement patterns for face recognition that are impervious to the influence of transitory mood changes (An & Hsiao, 2019) and able to predict both recognition performance and cognitive abilities, particularly executive and visual attention functions (Chan et al., 2018). These results suggest the possibility of using eye tracking for screening tests for cognitive deficits.

Here we trained the DNN+HMM model to perform face recognition and examined how its performance was associated with eye movement pattern and consistency at different learning stages. We expected that consistency and pattern of eye movements would better predict performance at early and late learning stages respectively. We then recruited children as participants and examined whether their face recognition performance was better predicted by consistency than pattern of eye movements.

Computational Modelling

Methods

Figure 2 shows the DNN+HMM model. An HMM generates a sequence of fixations, including location and spatial frequency (SF) scale to simulate attention window, according to its initial probabilities, transition matrix, and emission densities (assumed to be Gaussians). The attention window of each fixation is simulated by applying a Gaussian mask, centred on each fixation location/scale, to the input image. The masked images are fed into a multi-scale convolutional neural network (CNN) to extract image features at different SFs, which are then aggregated over time to form the visual short-term memory. At each time step, a multilayer perceptron (MLP) uses the current visual memory to predict the face class. Finally, the loss functions of the predictions across time steps are combined for training. During training, the HMM and CNN simultaneously learn the most appropriate sequence of fixations and perceptual representations from these fixations for face recognition.

We trained 80 models with different initializations, representing 80 individuals, using the aligned Labelled Faces in the Wild dataset (LFW-a, Wolf, Hassner, & Taigman, 2011). We selected the 100 most frequent people in the dataset (3,651 images), and used 90% of the data for training and 10% for validation. Each grayscale image was scaled to 64 pixels wide. Three SFs are used, 8, 16, and 32 cycles/face, which is the optimal SF range for face recognition (Costen, Parker, & Craw, 1996), with attention window sizes equivalent to 4° to 1° of visual angle (32 to 8 pixels) to simulate global/local attention. The attention window was simulated as a Gaussian mask centred at the fixation location (SD = half window size). We simulated saccade noise by adding Gaussian noise (SD = 0.375°, 3 pixels; Ohl, Brandt, & Kliegl, 2013) to each fixation. The SF attention process was implemented as a multi-scale CNN (for each SF, 2 layers of 3 x 3 filters with 8 and 16 channels) applied to an image pyramid of down-sampled images, with the original image used for

extracting high SF information and smaller images for low SF information.

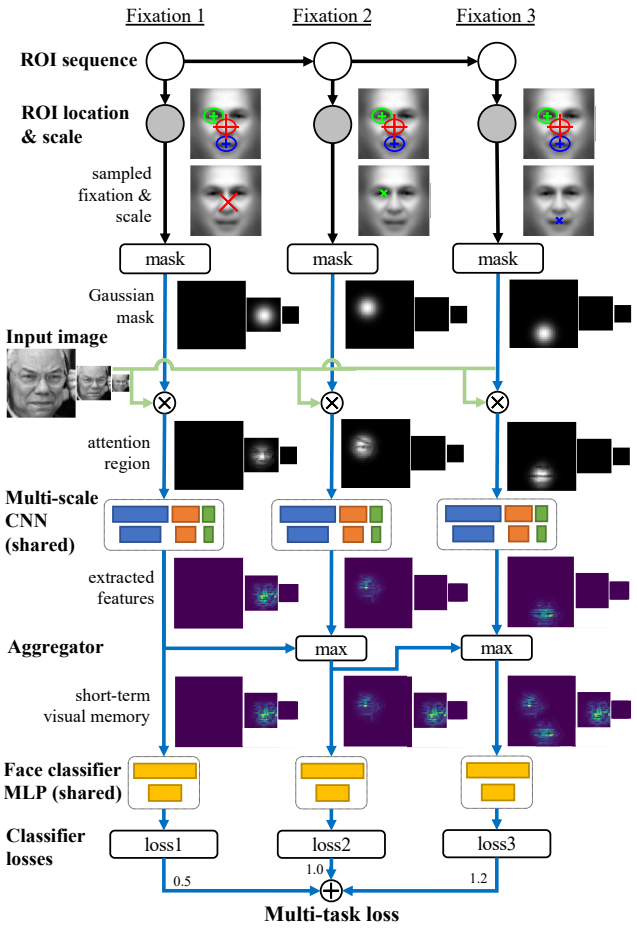


Figure 2: DNN+HMM for face recognition.

We assumed 3 fixations in sequence for recognition to match previous human subject studies (e.g., Hsiao & Cottrell, 2008; Chuk et al., 2014; Chan et al., 2018), as early fixations are shown to be more important for recognition (Hsiao & Cottrell, 2008; Chuk et al., 2017). For simplicity, we assumed a deterministic sequence for each individual. After the first fixation, each additional fixation updated an internal visual short-term memory representation by taking the maximum between the two feature maps. The internal representation was decoded into a face class using a shared MLP classifier (2 layers, 40 & 100 neurons). Thus, the model made 3 classifier predictions: from the first fixation, first 2 fixations, and all 3 fixations. Cross-entropy loss was applied to each prediction. The final loss was the weighted sum of these individual losses to make the first fixation the most informative, simulating face identification with as few fixations as possible. To encourage the model to move fixation locations towards informative features, the classification layers were regularized so that increasing the classifier weight on an informative feature has more penalty than moving a fixation towards the feature (which increases the strength of the image feature). Each convolution filter in the CNN was constrained to have unit norm. The fixation locations were also regularized to have a "centre" bias (Tatler, 2007). The model is initialized with 3 large ROIs with random location, and trained using the Adam optimizer (Kingma & Ba, 2014) for 500 epochs.

We assessed the models' eye movement behaviour after different numbers of training epochs. Eye movement consistency was assessed using the HMM's overall entropy (Cover & Thomas, 2006). Entropy is a measure of predictability: higher entropy indicates more random eye movements. Eye movement pattern was assessed using EMHMM. Specifically, all individual HMMs were clustered to discover two representative patterns A and B. Then, for each individual HMM, we defined AB scale as (A - B)/(|A| + |B|), where A and B referred to the model's data log-likelihood of pattern A and B respectively. Each model's similarity along the contrast between pattern A and B was quantified using AB scale (Chan et al., 2018).

Results

Figure 3 shows an example model after training. The ellipses represent the fixation ROIs (2 SD contours of the Gaussian emissions); the cross represents the attention window size, where larger windows correspond to using lower SF. This example model looks at the face centre using global attention (low SF), and then the eyes using local attention (medium SF). The CNN features selected in the MLP are visualized via the weight magnitudes of the first MLP layer, showing the use of global features on the eyes and nose, and local features on the eyes.

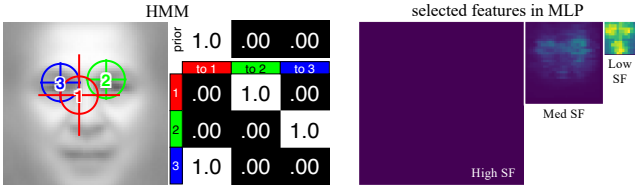


Figure 3: Example of a DNN+HMM after training.

We applied clustering to obtain two representative patterns for well-trained (adult) models at epoch 500, and two representative patterns for partially-trained (child) models at epoch 100 (Figure 4). The well-trained models exhibited an eyes-focused (pattern A) and a nose-focused strategy (pattern B), which differed significantly: the eyes-focused group's data log-likelihood given the eyes-focused HMM was higher than that given the nose-focused HMM, $t(36) = 15.82$, $p < .001$, $d = 2.60$; similarly for the nose-focused group, $t(42) = 7.25$, $p < .001$, $d = 1.11$. The partially-trained models exhibited similar strategies that also differed from each other (test on the eyes-focused group, $t(31) = 7.15$, $p < .001$, $d = 1.26$; test on the nose-focused group, $t(47) = 15.91$, $p < .001$, $d = 2.30$), albeit with much larger ROIs. In the well-trained models, those with pattern A (eye-focused) exhibited higher accuracy than pattern B, $t(78) = 3.33$, $p = .001$, d =

.75, M_A = .534, M_B = .500, and accuracy was positively correlated with AB scale, $p < 0.001$, R^2 = .161 (Figure 5 left). In contrast, the patterns A and B of partially-trained models did not differ in accuracy on the validation data, $t(78)$ = .72, p = .475, d = .16, M_A = .464, M_B = .457; the AB scale was not correlated with accuracy, R^2 = .161, p = .711. Thus, eye movement pattern was correlated with accuracy in well-trained models, but not partially-trained models.

Finally, we examined eye movement consistency (Figure 5 right). The entropy of partially-trained models was negatively correlated with accuracy, $p < .001$ R^2 = .205. In contrast, well-trained models did not show correlation between entropy and accuracy, as most well-trained models have converged to consistent, low entropy patterns.

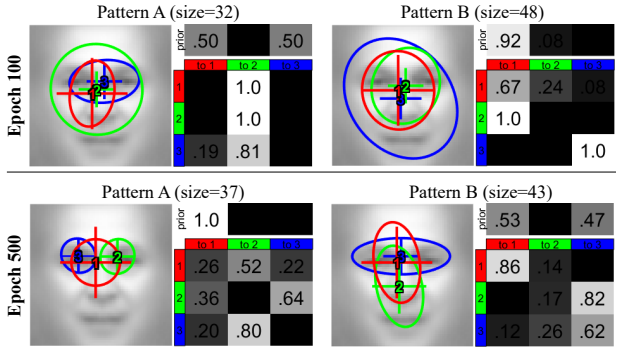


Figure 4: Representative HMMs for (top) partially trained (100 epochs) and (bottom) well-trained models (500 epochs).

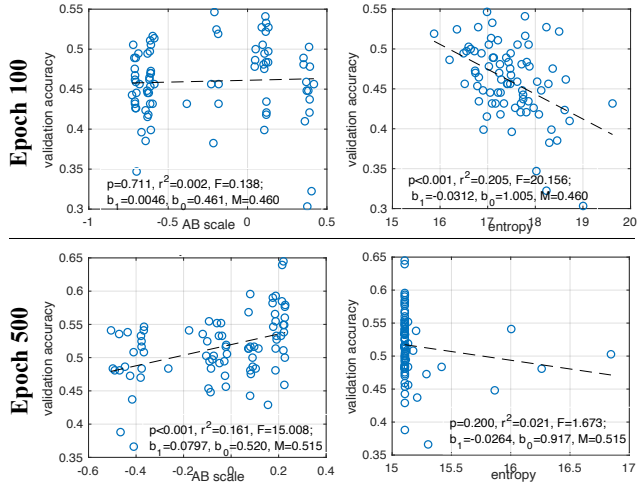


Figure 5: Entropy vs. AB scale and entropy vs. accuracy for (top) partially-trained and (bottom) well-trained models.

Experimental Study

Methods

Participants were 89 primary school students (40 females) from Hong Kong, aged 6 to 11 (M = 7.84). They had normal or corrected to normal vision and no cognitive deficits. They performed face recognition and cognitive ability tests with the order counterbalanced.

In the face recognition task, the stimuli consisted of 64 coloured frontal-view Asian adult face images with a neutral expression (half female). They were scaled and aligned to maintain the same inter-pupil distance, and cropped according to the face shape. The task consisted of two blocks, each with a study and a test phase. In the study phase, participants viewed 16 faces one at a time, each for 3 s, and were instructed to remember them. In the recognition phase, participants were presented with the 16 old and 16 new faces one at a time and asked to judge whether they saw the face in the study phase by a button response. The face was shown until response. Each trial began with a central fixation. A face was then presented either on the left or right of the screen (determined randomly). With a 60 cm viewing distance, the face spanned 8° of visual angle, and the face centre was 9° of visual angle away from the screen centre. Different images were used in the two blocks. Participants' eye movements were recorded using an SMI RED-n Scientific eye tracker (SensoMotoric Instruments GmbH). The right eye was tracked with 60 Hz sampling rate. A chinrest was used to minimize head movement. EMHMM was used to analyse eye movement data: each participant's data in the test phase was summarized into an HMM (see Chuk et al., 2017 for details). Following Chan et al. (2018), and also to match the modelling procedure, we used the first 3 fixations in each trial to train the HMM. Then we clustered all HMMs into two representative patterns, and calculated the AB scale as the modelling. In addition, we calculated each HMM's overall entropy as a measure of eye movement consistency.

The flanker task was used to measure selective attention ability (Eriksen & Eriksen, 1974). Each stimulus consisted of a target arrow pointing to either the right or left, two flanker arrows to the target's left, and two to the right. Congruent stimuli had flankers pointing in the same direction as the target; incongruent stimuli had flankers pointing in the opposite direction. Participants judged the target arrow direction by pressing keys. There were 120 trials. We measured the flanker effect as the performance difference between congruent and incongruent trials.

Spatial/verbal one-back tasks (Jaeggi, Buschkuehl, Perrig, & Meier, 2010) were used to assess working memory. In the spatial one-back task, in each trial, a blue square was presented at either above, to the right, to the left, or below a fixation cross for 500 ms, with an inter-trial interval of 2500 ms. Participants responded whether the square was at the same location as the previous trial by pressing buttons. There were 63 trials. The verbal one-back task had a similar procedure except that participants viewed a number presented at the screen centre instead of a blue square. D-prime and correct RT were measured.

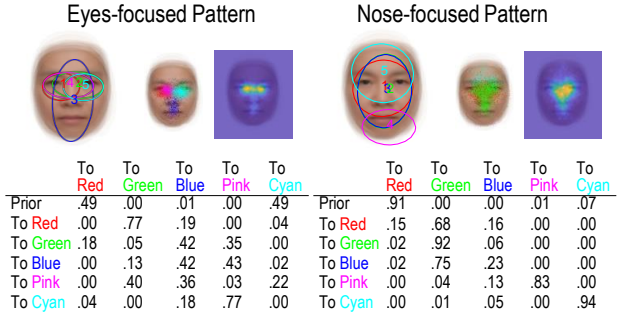
Trail making test (Reitan, 1958) was used to assess visual attention and task switching ability. In part A,

participants connected 25 circles from number 1 to 25 in a sequential order. In part B, participants connected numbers and alphabets alternatively in a sequential order. The tasks were given on two separate sheets of paper. The completion time and the number of errors were recorded.

Tower of London (TOL) test (Culbertson & Zillmer,1999) assessed planning and problem-solving abilities. Participants were presented with a start state (with 3 pegs and 3 beads) and a goal state and were instructed to move beads in the start state one at a time to reach the goal state using the least number of moves in 120 s. There were 10 trials with increasing difficulty. The percentage of completed trials, average number of excess moves, and correct RT were measured.

Results

Two children could not finish the face recognition and TOL task; one child could not finish the one-back task. EMHMM revealed two common patterns (Figure 6): the eyes-focused pattern (n = 37) mainly switched between the two eyes, with occasional fixations at the face centre, whereas the nose-focused pattern (n = 50) had more dispersed ROIs at the face centre. The two patterns differed significantly (test on the eyes-focused group; $t(36) = 16.04$, $p < .001$, $d = 2.64$; test on the nose-focused group, $t(49) = 4.89$, $p < .001$, $d = .69$). We quantified individual patterns' similarities along the eyes- and nose-focused pattern dimension using the EN scale $(E - N)/(|E| + |N|)$, where E and N stand for the data log-likelihood of the eyes- and nose-focused HMM respectively. ANCOVA was used on recognition performance D' with eye movement pattern as a between-subject variable and age as a covariate. Participants adopting the two patterns did not differ in performance, $F(1, 83) = .31$, $p = .58$. There was no correlation between EN scale and performance with age controlled, $r(83) = .12$, $p = .26$. (Figure 7a & c). In contrast, when we divided participants into high and low eye movement entropy groups, those with low entropy performed better, $F(1, 83) = 5.54$, $p = .021$, $\eta_p^2 = .063$, and entropy was correlated with performance with age controlled, $r(83) = -.29$, $p = 007$ (Figure 7b).



	To Red	To Green	To Blue	To Pink	To Cyan
Prior	.49	.00	.01	.00	.49
To Red	.00	.77	.19	.00	.04
To Green	.18	.05	.42	.35	.00
To Blue	.00	.13	.42	.43	.02
To Pink	.00	.40	.36	.03	.22
To Cyan	.04	.00	.18	.77	.00

	To Red	To Green	To Blue	To Pink	To Cyan
Prior	.91	.00	.00	.01	.07
To Red	.15	.68	.16	.00	.00
To Green	.02	.92	.06	.00	.00
To Blue	.02	.75	.23	.00	.00
To Pink	.00	.04	.13	.83	.00
To Cyan	.00	.01	.05	.00	.94

Figure 6. The eyes-focused (top) and nose-focused (bottom) representative strategies derived by clustering.

Among the cognitive ability measures, face recognition performance was correlated with the flanker effect in correct RT, $r(86) = .23$, $p = .030$. To examine whether eye movement entropy or flanker effect was a better predictor for recognition performance, a three-stage hierarchical regression was conducted with age and the flanker effect entered before eye movement entropy, and entropy was the only significant predictor, $\Delta R^2 = .205$, $F(1, 82) = 5.78$, $p = .018$ (Table 1). Tests for multicollinearity indicated a low level of multicollinearity among entered variables.

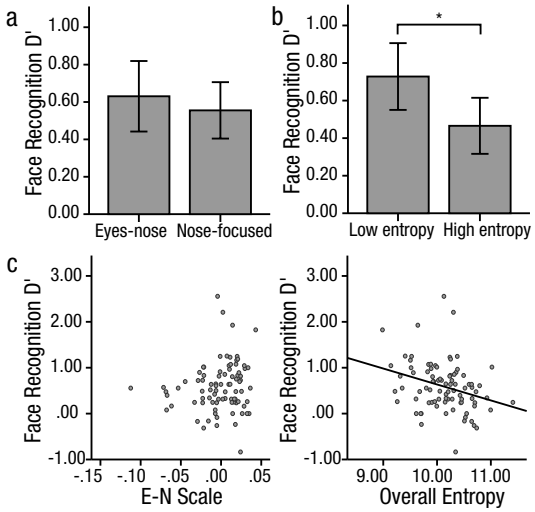


Figure 7: Recognition performance between (a) eyes- vs. nose-focused groups; (b) low vs. high entropy groups (Error bars are 95% CIs. * $p < 0.05$) (c) Recognition performance was correlated with overall eye movement entropy, but not eye movement pattern (EN scale).

Table 1. Summary of hierarchical regression analysis

	β	t	R	R^2	ΔR^2
Step1			.19	.036	.036
Age	.19	1.78			
Step 2			.29	.086	.049*
Age	.17	1.56			
Flanker effect in correct RT	.22	2.12*			
Step 3			.38	.15	.060*
Age	.15	1.47			
Flanker effect in correct RT	.17	1.66			
Entropy	-.25	-2.40*			

Note. *p < .05

We then examined what cognitive abilities best accounted for consistency and pattern of children's eye movements through stepwise hierarchical multiple regression analysis with age entered as a covariate at the first step. Tests indicated a low level of multicollinearity among the variables. Eye movement pattern/EN scale was best predicted by verbal one-back D', $R^2 = .075$, $F(1, 79) = 3.19$, $p = .047$. In contrast, eye movement entropy was best predicted by flanker effect in correct RT, $\beta = -.29$, $t = -2.67$, $p = .009$, and TOL % of completed trials, $\beta = .25$, $t = 2.36$, $p = .021$; $R^2 = .013$, $F(3, 78) = 3.75$, $p = .014$.

Thus, eye movement pattern may be related to working memory, whereas eye movement consistency was associated with selective attention and executive function.

Discussion

In perceptual expertise research, it has long been assumed that looking at diagnostic features leads to better performance (Chuk et al., 2017). However, typically we can only attend to features one at a time, and this requires eye movement planning. Recent research has shown that people exhibit person-specific eye movement patterns in face recognition, and deviation from such visual routines can impair performance (Peterson & Eckstein, 2013). While it suggests the importance of visual routines, how eye movement consistency contributes to performance has been overlooked in the literature. Here we aimed to fill this gap by testing the hypothesis that eye movement consistency plays an important role during early learning, since inconsistent eye fixations during early learning, even if landing on diagnostic regions, can signify difficulty in discovering and extracting diagnostic features to develop a visual routine, leading to suboptimal performance. Also, once a suboptimal routine is formed, it can become difficult to change. Indeed, adult face recognition performance has limited plasticity for improvement through training (Tree et al., 2017), and this phenomenon may be related to a stable visual routine.

Consistent with our hypothesis, our DNN+HMM model trained for face recognition showed that during early training, the model's performance was well predicted by consistency but not pattern of eye movements. In contrast, in fully trained models, eye movement pattern predicted performance whereas eye movement consistency did not. Similarly, in our human data, children's recognition performance was well predicted by eye movement consistency with age and cognitive abilities controlled, but not by eye movement pattern. Previous studies with adult participants have shown that eye movement pattern was predictive of recognition performance (Chuk et al., 2017; Chan et al., 2018). We reanalysed the data to examine the effect of eye movement consistency. Stepwise multiple regression analysis using data from Chuk et al. (2017) showed that recognition performance was best predicted by eye movement pattern, $R^2 = 21.8\%$, $p = .001$, and adding consistency did not significantly account for additional variance, $\Delta R^2 = .002$, $p = .726$. Similarly, using data from Chan et al. (2018), recognition performance was correlated with eye movement pattern, $r = -.25$, $p = .041$, but not consistency. Thus, eye movement consistency predicts early learning performance, whereas eye movement pattern predicts late, expert-level performance.

The modelling data suggested that during early learning, some models may have fixations at diagnostic features with an inappropriate attention window size, resulting in suboptimal performance and lower likelihood of selecting the same location. Thus, at this stage, fixations on diagnostic features did not reflect better performance. In contrast, once an optimal fixation location and attention window size combination was selected, it was likely to be selected again, leading to more consistent eye movements. Models with difficulty discovering optimal feature location and attention window size combinations might end up with a suboptimal eye movement pattern. Thus, when this process continued and all models converged to a similar level of eye movement consistency, performance became better predicted by eye movement pattern. In addition, in our human data, children's eye movement consistency was best predicted by executive function abilities; similarly, adults' eye movement pattern was associated with executive function abilities (Chan et al., 2018). These findings suggest that executive function abilities may underlie this learning process, affecting eye movement consistency in children and being reflected in eye movement pattern in adults. It also suggests that deficits in executive functions may underlie face recognition difficulties. Children with recognition difficulties, such as autism, may benefit from training to more efficiently develop a consistent eye movement pattern that follows the optimal strategy. Future work will examine these possibilities.

In the current modelling, to focus our examinations on the interaction between DNN and HMM during learning, the transition matrix and prior were fixed so that the ROI sequence was deterministic. Future work may allow the transition matrix and prior to be learned by using Gumbel-Softmax reparameterization (Jang, Gu, & Poole, 2017) of these probability models so that the ROI sequences will be generated according to the prior and transition matrix. Also, bottom-up attention may be simulated by learning a conditional transition matrix that takes previously attended features as input, so that attention can be guided by both bottom-up and top-down information. We may train a face-space embedding using triplet loss (Wang & Deng, 2019) to facilitate generalization with a large number of face labels. These may enhance the model's cognitive plausibility in future studies.

In conclusion, through computational and experimental examinations, we show that learning to recognise faces initially involves developing a consistent visual routine, which depends on executive function abilities. As the result, children's face recognition performance is better predicted by eye movement consistency rather than eye movement pattern, in contrast to adult face recognition. These findings have significant implications for ways to enhance learning in both healthy and clinical populations. The proposed computational model can be applied to other learning tasks, further increasing the impact.

Acknowledgments

We are grateful to RGC of Hong Kong (Project # 17609117 to Hsiao). We thank Calise Po Tik Lau and Yueyuan Zheng for their help in reanalysing data from previous studies.

References

An, J. H., & Hsiao, J. H. (in press). Modulation of mood on eye movement and face recognition performance. *Emotion*.

Ablavatski, A., Lu, S., & Cai, J. (2017). Enriched deep recurrent visual attention model for multiple object recognition. *Applications of Computer Vision*, 971–978.

Barrington, L., Marks, T., Hsiao, J. H., & Cottrell, G. W. (2008). NIMBLE: A kernel density model of saccade-based visual memory. *J. Vis.*, *8(14):7*, 1-14.

Chan, C. Y. H., Chan, A. B., Lee, T. M. C., & Hsiao, J. H. (2018). Eye-movement patterns in face recognition are associated with cognitive decline in older adults. *Psychol. B. Rev.*, *25*(6), 2200–2207.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *J. Vis.*, *14*(11):8, 1-14.

Chuk, T., Crookes, K., Hayward, W. G., Chan, A. B., & Hsiao, J. H. (2017). Hidden Markov model analysis reveals the advantage of analytic eye movement patterns in face recognition across cultures. *Cognition*, *169*, 102-117.

Costen, N. P., Parker, D. M., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Percept. Psychophys.*, *58*(4), 602–612.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons.

Coviello, E., Chan, A.B., & Lanckriet, G.R.G. (2014). Clustering hidden Markov models with variational HEM. *J. Mach. Learn. Res.*, *15*(Feb), 697-747.

Culbertson, W. C., and Zillmer, E. A. (1999). *The Tower of London, Drexel University, Research Version: Examiner's Manual*. New York: Multi-Health Systems.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Percept. Psychophys., 16*(1), 143-149.

Hsiao, J. H., & Cottrell, G. W. (2008). Two fixations suffice in face recognition. *Psychol. Sci.*, *9(10)*, 998-1006.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394-412.

Jang, E., Gu, S., & Poole, B. (2017). Categorical Reparameterization with Gumbel- Softmax. *Intl. Conf. on Learning Representations (ICLR)*.

Kingma, D. K., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. https://arxiv.org/abs/1412.6980

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems*, 2204-2212.

Nazir, T. A., & O'Regan, J. K. (1990). Some results on the translation invariance in the human visual system. *Spatial Vision*, *3*, 81–100.

Ohl, S., Brandt, S. A., & Kliegl., R. (2013). The generation of secondary saccades without postsaccadic visual feedback. *J. Vis.*, *13*(5):11, 1–13

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

Reitan, R. M. (1958). The validity of the Trail Making Test as an indicator of organic brain damage. *Percept. Mot. Skills*, *8*, 271-276.

Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vision*, *7*(14): 4.

Tree, J. J., Horry, R., Riley, H., & Wilmer, J. (2017). Are portrait artists superior face recognizers? Limited impact of adult experience on face recognition ability. *J. Exp. Psychol. Hum. Percept. Perform.*, *43*(4), 667-676.

Vinette, C., Gosselin, F., Schyns, P. G. (2004). Spatiotemporal dynamics of face recognition in a flash: it's in the eyes. *Cogn. Sci.*, *28*, 289–301.

Wang, M., & Deng, W. (2019). Deep face recognition: A survey. *arXiv*. https://arxiv.org/abs/1804.06655

Wilson, C. E., Palermo, R., Brock, J. (2012) Visual scan paths and recognition of facial identity in autism spectrum disorder and typical development. *PLoS ONE*, *7*(5): e37681.

Wolf, L., Hassner, T., & Taigman, Y. (2011). Effective Face Recognition by Combining Multiple Descriptors and Learned Background Statistics. *IEEE Trans. Pattern Anal. Mach. Intell.*, *33*(10), 1978–1990.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA.*, *111*(23), 8619-8624.