Is Holistic Processing Associated with Face Scanning Pattern and Performance in Face Recognition? Evidence from Deep Neural Network with Hidden Markov Modeling

Wei Xing[#] (psywxing@connect.hku.hk) Department of Psychology, University of Hong Kong

Yueyuan Zheng[#] (mercuryzheng@connect.hku.hk) Department of Psychology, University of Hong Kong

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

Janet H. Hsiao (jhhsiao@ust.hk) Division of Social Science, Hong Kong University of Science and Technology

Abstract

Here we used deep neural network + hidden Markov model (DNN+HMM) to provide a computational account for the relationship among holistic processing (HP), face scanning pattern and face recognition performance. The model accounted for the positive associations between HP and eyes-focused face scanning pattern/face recognition performance observed in the literature regardless of the version of the composite task used to measure HP. Interestingly, we observed a quadratic relationship between HP and face scanning pattern, where models being highly eyes-focused or highly nose-focused had lower HP. By inspecting fixation locations and associated attention window size in the model and XAI methods, we found that the eyes- and nose-focused models both developed local and holistic internal representations during training, and their difference was in the temporal dynamics of how these representations were used. Our findings demonstrated how computational modeling could unravel the mechanisms underlying cognition not readily observable in human data.

Keywords: holistic processing; face scanning pattern; face recognition; computational modeling

Introduction

The ability to recognize faces is of vital importance in daily life. Previous studies have reported that holistic processing (HP), the phenomenon that people perceive a face as an inseparable whole rather than separate components (Rossion, 2013), marks perceptual expertise of faces (Richler & Gauthier, 2014). In addition, eye-focused face scanning behavior was found to be advantageous to face recognition performance (Hsiao et al., 2022a). However, the relationship among HP, face scanning pattern and face recognition performance remains inconclusive. Here we aimed to provide a computational account on the relationship among HP, face scanning pattern and face recognition performance.

Previous research has measured HP using different task paradigms (e.g., composite face, part-whole, and inverted face paradigms) and designs (e.g., standard and partial designs of composite face paradigm). For example, using the composite face paradigm, two identical top-halves of a face can be perceived as different when they are combined with two different bottom halves (Hole, 1994). HP measured using this paradigm thus reflects the extent to which people can neglect an irrelevant part of a face (Young et al., 1987). Although HP is shown to increase with years of experience with faces during development, whether it is associated with face recognition performance remains inconclusive (Rossion et al., 2013), since different task designs of the paradigm have led to inconsistent findings. For example, using the complete design where the *congruency effect* was used to measure HP, higher HP was associated with better face recognition performance (DeGutis et al., 2013). However, this correlation was not always observed using the standard design where the *alignment effect* was used to measure HP (Rezlescu et al., 2017). This inconsistency may result from different aspects of HP being measured: It has been argued that the standard design measures perceptual integration (Rossion, 2013), whereas the complete design measures failure of selective attention (Richler et al., 2008). However, it remains unclear whether these measures indeed reflect different mechanisms and inconclusive whether HP measured using this paradigm is associated with face recognition performance.

Face scanning pattern plays a vital role in face recognition. Since eyes are the most diagnostic features for recognizing a face (e.g., Gosselin & Schyns, 2001), individuals who attend more to the eyes during face recognition have better face recognition performance. This phenomenon has been observed in studies either using predefined ROI (e.g., Davis et al., 2017) or data-driven approaches in eye movement data analysis (e.g., Chan et al., 2018). This eyes-focused face scanning pattern was shown to be associated with the use of high spatial frequency (SF) information (local attention) during face identification (Miellet et al., 2011). Indeed, local attention priming using Navon stimuli led to more eyes-focused face scanning pattern and enhanced performance in face recognition as compared with global priming (Cheng et al., 2018). In contrast, global attention priming enhanced HP of faces using the complete composite paradigm (Gao et al., 2011). These results suggested that a more eyes-focused face scanning pattern is associated with the engagement of local attention, which should in turn lead to reduced HP. Consistent with this speculation, a recent study observed that a more eyes-focused face scanning pattern was associated with lower HP using the part-whole paradigm (Hsiao et al., 2021a).

However, inconsistent with this speculation, Zhong et al. (2024) have recently found that individuals who adopted a more eye-focused face scanning pattern showed stronger HP (as measured using the complete composite paradigm) than those being nose-focused. Indeed, as both eyes-focused face

[#] These authors contributed equally to this study, and they are considered as co-first authors.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

scanning pattern and HP have been shown to be beneficial to face recognition performance, a more eyes-focused pattern may be associated with stronger HP. Thus, the relationship between HP and face scanning pattern remains unclear.

Accordingly, here we aimed to address these research gaps through computational modeling. Specifically, we aimed to provide computational explanations for (1) the relationships among HP, face scanning pattern and face recognition performance, and (2) whether the results differed when HP was measured using the standard vs. the complete composite paradigms. In the literature, neural network models have been commonly used to simulate human face recognition and account for face processing effects including HP (Omigbodun & Cottrell, 2013; Hsiao & Galmar, 2016). Nevertheless, earlier models did not typically take human eye movements and the associated attention windows into account. To better understand the relationship between eye movement pattern and perceptual representation development during learning to recognize faces, Hsiao et al. (2022) proposed deep neural network + hidden Markov model (DNN+HMM), where the DNN learns the optimal perceptual representations under the guidance of an attention mechanism summarized in an HMM, and the HMM learns the optimal face scanning pattern through the feedback from the DNN. This model was able to account for the relationship between face scanning pattern and face recognition performance in human data: the eyes-focused pattern predicted better face recognition in adults (Hsiao et al., 2021a), whereas higher eye movement consistency predicted better face recognition in children (Hsiao et al., 2022). Thus, DNN+HMM has good cognitive plausibility in modeling face scanning pattern learning in face recognition. Here we adopted DNN+HMM to examine the relationship among HP, face scanning pattern, and face recognition performance. We expected that the model would be able to account for the association between HP and face scanning pattern or face recognition performance observed in the literature regardless of whether the standard or the complete composite design was used to measure HP. In addition, through examining the differences in fixation location, associated attention window size, and internal representation between models with different face scanning patterns, we expected the modeling to provide computational explanations on scenarios where HP may be positively or negatively correlated with face scanning pattern.

Method

DNN+HMM Model Structure and Training

We adopted the 80 well-trained DNN+HMM models for face recognition with the same configuration and model weights as Hsiao et al. (2022). The DNN+HMM models (Figure 1) first generated three fixation locations and their corresponding SF scale based on its HMM's initial probabilities, transition matrix, and emission density (assumed to follow a Gaussian distribution). The SF scale was designed to simulate different attention window sizes. To extract features at these three scales, we generated a set of multi-scale images, in the size of 64×64, 32×32, 16×16, corresponding to high SF features (smallest attention window), middle SF features, and low SF features (largest attention window), respectively. For each fixation, a mask was generated using a Gaussian emission at the fixation location and applied to the corresponding multi-scale input image. The masked images were fed into convolutional sub-networks to extract image features at each SF scale. Each sub-network consisted of two convolutional blocks with 8 and 16 output channels respectively, using 3×3 convolution kernels and ReLU activation function. The image features were aggregated across fixations using element-wise maximum, thereby simulating visual short-term memory. A multi-layer perceptron then used current visual memory to predict the face categories, which consisted of two fully connected layers: a 40-neuron hidden layer (with ReLU activations) and a 100-neuron output layer with a SoftMax activation function to produce class probabilities. The parameters shared across fixations. The prediction was supervised using categorical cross entropy loss, and the total loss was the weighted sums over fixations. After training, the DNN+HMM acquired face scanning patterns and perceptual representations that were favorable for face recognition tasks.

The models were trained on a subset of the LFW-a dataset (Wolf et al., 2011), comprising aligned faces. The 100 most frequent individuals were selected to form a dataset of 3,651 images, 90% of which were randomly selected as training set and the rest as validation set. Each model was trained for 500 epochs to ensure that the model is well-converged, and different weight initializations yielded different models.

Design

The HP effect was examined at three levels, following previous studies (Hsiao & Galmar, 2016; Omigbodun & Cottrell, 2013): the early perceptual representation from the last convolutional (LastConv) layer, the intermediate representation from the fully-connected (FC) layer, and the face identity representation from the output layer, which was typically measured in a human behavioral task of HP. For each concerned layer of models, a paired sample t-test between aligned and misaligned conditions for the standard design and a 2 (congruency: congruent vs incongruent) × 2 (alignment: aligned vs misaligned) ANOVA for the complete design was conducted for HP task accuracy to examine whether models exhibited HP. In addition, for each layer, we did the curve estimation to examine the relationship between HP and face scanning pattern/face recognition performance.

Face Recognition Task

Face recognition task was used to assess face recognition ability of the DNN+HMM models. The stimuli included the 366 face images (the validation set of LFW-a dataset). Models judged the identity of the face images, which were learnt during training. We measured the performance of models as the accuracy of predictions for the validation set.

Data Analysis for Models' Fixation Behavior

A variational hierarchical expectation maximization algorithm (Coviello et al., 2014) was applied to cluster the indi-

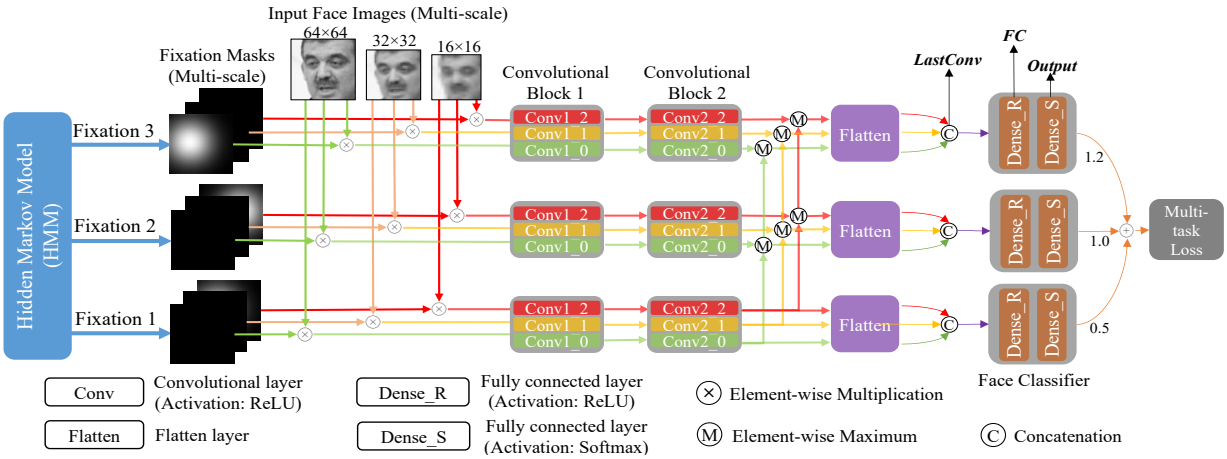


Figure 1. The DNN+HMM for face recognition. The bold-italic highlighted text shows the layers where we extracted representation features in the DNN+HMM model: ***LastConv, FC40***, and ***Output layers***. Note that the extraction was only conducted in the time step of the last fixation, for its aggregation of the representation from all fixations.

vidual HMMs of the 80 models and generate two representative fixation patterns (HMMs). Following previous studies (Hsiao et al., 2021b; Zheng & Hsiao, 2023; Zheng et al., 2022), models' fixation pattern was quantified using A-B scale, which was calculated as $\frac{(LA - LB)}{(|LA| + |LB|)}$, where LA and LB represent log-likelihoods of a model's HMM being generated by the representative HMMs A and B respectively. A higher A-B scale indicated higher similarity to Pattern A.

Model Visualization using Explainable AI

We used eXplainable AI (XAI) method GradCAM (Selvaraju et al., 2017), one of the most commonly used visualization techniques in XAI, to visualize which of the model's internal representations (features) or image regions contributed the most to the face identification. GradCAM produces a saliency map from the feature map by weighting the feature channels by their average gradient, which measures the influence of the feature on the output prediction. Specifically, we used GradCam to visualize: (1) for each SF scale, the features used for the output prediction at the LastConv layer, averaged over the individual models in a fixation pattern group; and (2) features used by each active neuron in the FC layer, which explain the types of internal representations used in the model.

Composite Face Task

Composite Face Task (Gauthier & Bukach, 2007) was used to assess HP of DNN+HMM models. The models were tested to see if representations of the attended upper halves of two faces were the same or different (Chung et al., 2018). We first randomly selected 100 face images from 100 different celebrities in the validation set of LFW-a dataset. We then generated 500 different pairs of face images by randomly drawing from all possible combinations of the face images. For each original image pair, we created four manipulated image pairs based on congruency by response (same vs different) conditions (Figure 2) for each alignment condition. In the congruent trials, the top halves and bottom halves of the two faces elicit the same response. In the incongruent trials, the two halves lead to different responses. Each image of the image pairs was center cropped to the resolution of 225×225, resized and generated multi-scale face images as the input of the models. To simulate attention to the upper half of the face, following previous studies (Hsiao & Galmar, 2016), the pixel value of the lower half of all images was attenuated by multiplying a factor of 0.5. To simulate misaligned trials, we masked the lower half of the face by multiplying the pixel value with a factor 0 (i.e., no attention).

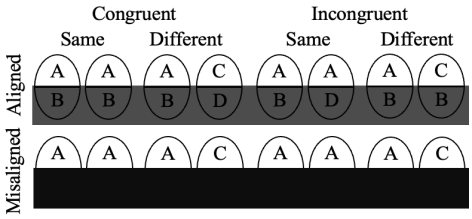


Figure 2. Face stimuli of the composite face task. Note that the stimuli were presented in pairs.

We calculated the HP effect at the three stages from the last fixation (after aggregating across previous fixations). For the LastConv and FC layers, we followed the approach from Hsiao and Galmar (2016). In each task trial, the output representations of each stimulus were flattened into a one-dimensional vector, and the correlation between the two vectors was calculated as a measure of similarity. The model's response was regarded as "same" if the correlation of a stimulus pair was above a threshold, and as "different" if it was below the threshold. The threshold was set as the midpoint between the averaged correlation value of the "same" stimulus pairs and that of the "different" pairs. For the output layer, we compared the output identities of the two input faces (Omigbodun & Cottrell, 2013). If the output identities of the two images in a trial were the same, the response was regarded as "same", otherwise the response was "different". For the standard design, HP (i.e., alignment effect only) was calculated as (*MI* - *AI*), where *MI* and *AI* denoted accuracy of misaligned and aligned incongruent conditions. For the complete design, HP

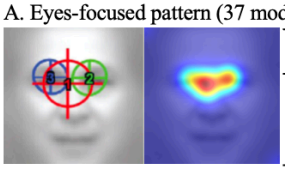
(i.e., interaction between congruency and alignment) was calculated as ((*AC* - *AI*) - (*MC* - *MI*)), where *AC, AI, MC* and *MI* denoted d' of the four alignments by congruency conditions.[1]

Results

Face Scanning Pattern during Face Recognition

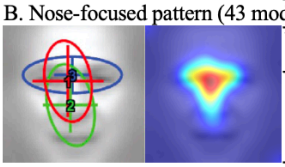
Here we discovered two representative face scanning patterns (see also Chuk et al., 2014; Hsiao et al., 2022; Figure 3): eyes-focused and nose-focused patterns. After the first fixation at the face center/red ROI (100%) with low SF scale (large attention window size), models adopting the eyes-focused pattern typically started to fixate on the two eyes, including green (52%) and blue (22%) ROIs, with middle SF scale (medium window size). In contrast, models with the nose-focused pattern started at the face center (red ROI: 53%) with low SF scale or the general eye region (blue ROI: 47%) with middle SF scale. They mainly looked at the same region afterwards and sometimes shifted to green ROI covering the nose and mouth region. The two representative patterns differed significantly from each other: data from models using the eyes-focused pattern were more likely to be generated from the eyes-focused than nose-focused HMM, $t(36) = 15.82$, $p < .001$; vice versa for the nose-focused pattern, $t(42) = 7.25$, $p < .001$. Here the A-B scale was referred to as the Eyes-Nose scale (EN scale). The EN scale was positively correlated with face recognition accuracy, $r(78) = .40$, $p < .001$, suggesting that more eyes-focused models performed better.

A. Eyes-focused pattern (37 models)



Eyes-focused	To Red	To Green	To Blue
Priors	1.0	.00	.00
From Red	.26	.52	.22
From Green	.36	.00	.64
From Blue	.20	.80	.00

B. Nose-focused pattern (43 models)



Nose-focused	To Red	To Green	To Blue
Priors	.53	.00	.47
From Red	.86	.14	.00
From Green	.01	.17	.82
From Blue	.12	.26	.62

Figure 3. The (A) eyes-focused and (B) nose-focused patterns from 80 models. Ellipses show ROIs as 2-D Gaussian emissions. The crosses at the center of ROIs show the associated attention window size. The table shows transition probabilities among the ROIs. Priors show the probabilities that a fixation sequence starts from the ellipse. The image in the middle shows the corresponding heatmap.

XAI Saliency Maps for LastConv and FC Layers

The XAI saliency maps showed that at the LastConv layer, models from both eyes-focused and nose-focused groups involved local attention on middle SF scale around the eye region and global attention on low SF scale (Figure 4A). At the FC layer, different active neurons focused on different features with different attention window sizes. More specifically, in both eyes-focused and nose-focused models, some active nodes focused on the specific features (e.g., eyes and mouth) with local attention, while some other nodes attended the whole face with global attention (Figure 4B).

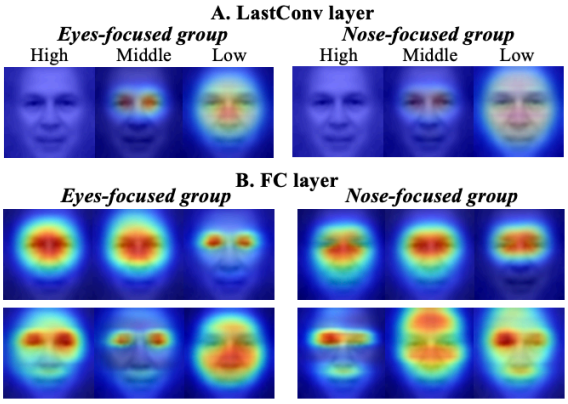


Figure 4. (A) XAI saliency maps for the LastConv layer averaged across trials and models for eyes-focused and nose-focused groups under high, middle, and low SF scales. (B) The active nodes with the highest three weightings of the FC layer from 2 models in eyes- and nose-focused groups.

Holistic Processing at the LastConv Layer

Standard Design of Holistic Processing

At the LastConv layer, models had better accuracy for the misaligned than aligned trials, $t(79) = 6.55$, $p < .001$, $d = .73$, suggesting that they exhibited HP. HP was not correlated with face recognition accuracy or EN scale, $ps > .05$. There was no quadratic relationship between HP and EN scale, $R^2 = 0.04$, $F(2, 77) = 1.65$, $p = .200$ (Figure 5A).

Complete Design of Holistic Processing

There was a main effect of alignment, $F(1, 79) = 336.08$, $p < .001$, $\eta^2_p = .81$, a main effect of congruency, $F(1, 79) = 445.16$, $p < .001$, $\eta^2_p = .85$, and an interaction between them, $F(1, 79) = 449.82$, $p < .001$, $\eta^2_p = .85$: models performed better in the congruent than incongruent trials only in the aligned condition, $t(79) = 21.25$, $p < .001$, but not in the misaligned condition, $p = .987$, suggesting that the models exhibited HP. HP was not correlated with face recognition accuracy or EN scale, $ps > .05$. Interestingly, a quadratic relationship between HP and EN scale, $R^2 = 0.13$, $F(2, 77) = 5.59$, $p = .005$, $\beta1 = -0.99$, $\beta2 = -4.09$ (Figure 5D) was found. This result indicated that models adopting a highly eyes-focused or a highly nose-focused pattern tended to have lower HP, while those using a mixture of the two patterns tended to have higher HP.

Holistic Processing at the FC Layer

Standard Design of Holistic Processing

Models had better accuracy for the misaligned than aligned trials, $t(79) = 15.36$, $p < .001$, $d = 1.72$, suggesting that they

[1] We also calculated unnormalized HP measure, (*AC-AI*), and the results were consistent with those using the normalized measure.

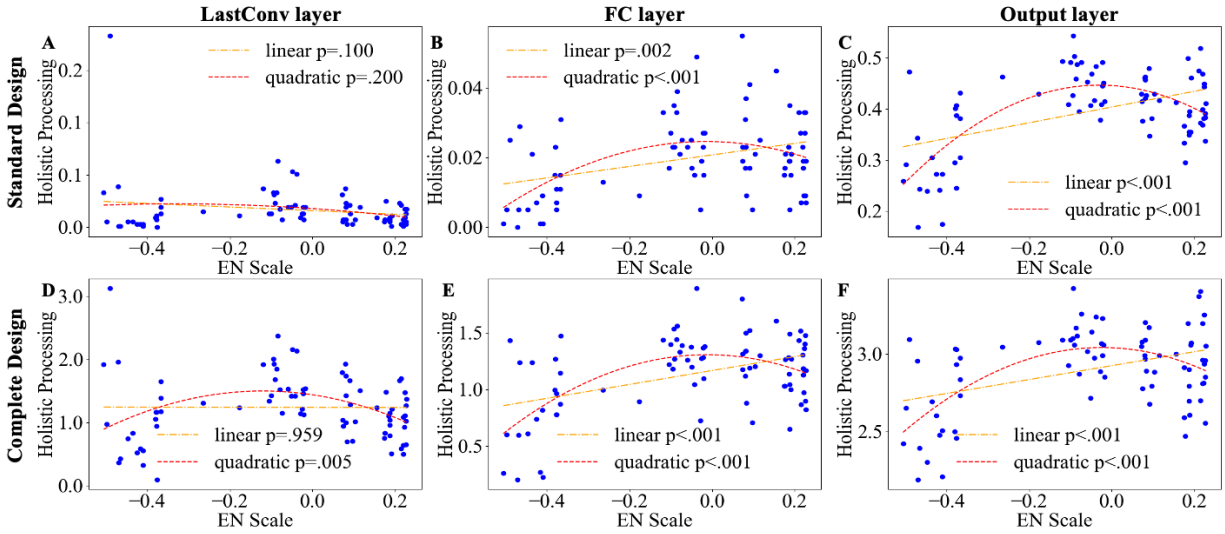


Figure 5. The linear and quadratic relationship between HP and EN scale in each concerned layer.

exhibited HP at FC layer. HP showed a linear, $r(78) = .35$, $p < .001$, as well as a quadratic relationship with EN scale, $R^2 = 0.22$, $F(2, 77) = 10.69$, $p < .001$, $\beta1 = -0.002$, $\beta2 = -0.079$ (Figure 5B). However, HP was not associated with face recognition accuracy, $r(78) = .02$, $p = .865$.

Complete Design of Holistic Processing

There was a main effect of alignment, $F(1, 79) = 429.61$, $p < .001$, $\eta^2_p = .85$, a significant main effect of congruency, $F(1, 79) = 819.00$, $p < .001$, $\eta^2_p = .91$, and an interaction between them, $F(1, 79) = 873.39$, $p < .001$, $\eta^2_p = .92$: models had better performance in the congruent than incongruent trials only in the aligned condition, $t(79) = 29.39$, $p < .001$, but not in the misaligned condition, $p = .972$. This indicated that the models exhibited HP. HP had both a linear, $ps < .001$, and a quadratic relationship with EN scale, $R^2 = 0.33$, $F(2, 77) = 19.34$, $p < .001$, $\beta1 = -0.07$, $\beta2 = -2.86$ (Figure 5E), but was not correlated with recognition accuracy, $p > .05$.

Holistic Processing at the Output Layer

Standard Design of Holistic Processing

Models had better accuracy for the misaligned than aligned trials, $t(79) = 115.18$, $p < .001$, $d = 12.88$, suggesting that models exhibited HP at output layer. HP had both a linear, $r(78) = .47$, $p < .001$, and a quadratic relationship with EN scale, $R^2 = 0.46$, $F(2,77) = 33.12$, $p < .001$, $\beta1 = -0.05$, $\beta2 = -0.86$ (Figure 5C). In addition, HP was positively correlated with face recognition accuracy, $r(78) = .26$, $p = .021$.

Complete Design of Holistic Processing

There was a main effect of alignment, $F(1, 79) = 1037.26$, $p < .001$, $\eta^2_p = .93$, a significant main effect of congruency, $F(1, 79) = 7775.75$, $p < .001$, $\eta^2_p = .99$, and an interaction between them, $F(1, 79) = 9242.47$, $p < .001$, $\eta^2_p = .99$: models performed better in the congruent than incongruent trials in the aligned condition, $t(79) = 93.91$, $p < .001$, but not in the misaligned condition, $p = .993$. This indicated that the models showed HP. HP had both a linear, $r(78) = .40$, $p < .001$, and a quadratic relationship with EN scale, $R^2 = 0.32$, $F(2,77) = 18.39$, $p < .001$, $\beta1 = -0.12$, $\beta2 = -2.36$ (Figure 5F). HP was positively face recognition accuracy, $r(78) = .25$, $p = .028$.

Discussion

Here we examined the relationship among holistic processing (HP), face scanning pattern, and face recognition performance through computational modeling. The results were summarized in Table 1. We found that models showed an HP effect at all processing stages/layers regardless of whether HP was measured using either the standard or the complete design. The models also showed a positive correlation, as well as a quadratic relationship, between HP and eyes-focused face scanning pattern using the HP measures from the intermediate representation at the FC layer and the face identity representation at the output layer. Specifically, models that showed highly eyes-focused or highly nose-focused patterns tended to have lower HP, while those whose face scanning patterns were a mixture of the two patterns tended to have higher HP. Finally, HP as measured from the face identification stage was positively correlated with face recognition performance. In general, the results were consistent regardless of whether the standard design or the complete design was adopted to measure HP.

Our results showed that greater HP was associated with better face recognition performance when HP was measured using the face identity representation at the output layer. This finding was generally consistent with the previous studies using the complete design (Richler et al., 2011), though this association was not always observed using the standard design (Rezlescu et al., 2017). This inconsistency may result from a lower testing power of the standard design (Richler & Gauthier, 2014). After enabling a greater testing power with a larger number of trials (500 trials per condition) in the current study, the positive relationship between HP and face recognition performance was consistently found using both designs. Note that at the output layer, HP was measured according to whether the model identified the two input faces (with the bottom halves attenuated) as the same face identity. Since face recognition performance was also measured from

	LastConv layer		FC layer		Output layer	
	Standard	Complete	Standard	Complete	Standard	Complete
Is there HP?	✓	✓	✓	✓	✓	✓
Is HP positively correlated with EN scale?	×	×	✓	✓	✓	✓
Is there a quadratic relationship between HP and EN scale?	×	✓	✓	✓	✓	✓
Is HP positively correlated with face recognition performance?	×	×	×	×	✓	✓

Table 1. An overview of results comparisons across different layers and composite task paradigms of the relationship among HP, face scanning pattern (EN scale), and face recognition performance. ✓and × denote yes and no respectively.

the output layer, this may explain why the correlation with face recognition performance was only observed when we measured HP at the output layer. This result suggests that the positive association between HP and face recognition performance observed in human data may originate from the face identification processes of the composite face task and face recognition task.

However, HP may not always predict better recognition performance. Previous studies reported that people with face drawing experience showed lower HP than novices but did not differ in face recognition performance from novices (Zhou et al., 2012). This suggested that being more or less holistic due to better perceptual expertise may not be associated with face recognition performance. Consistent with these findings, here we found that HP was not associated with face recognition performance if it was measured using the representations at the early perceptual and intermediate stages. Thus, whether HP is associated with face recognition performance may depend on whether participants' responses in the composite face task rely more on the matching of face identities or lower-level features.

In addition, we found that more eyes-focused models showed greater HP when HP was measured at the intermediate or the late face identification stage. This result is consistent with a recent study where people who adopted a more eyes-focused face scanning pattern showed greater HP (Zhong et al., 2024). Interestingly, we observed a quadratic relationship between HP and face scanning pattern, suggesting that models adopting a mixture of eyes-focused and nose-focused patterns have greater HP, while those being highly eyes-focused or highly nose-focused have lower HP. By inspecting the representative eyes- and nose-focused patterns in Figure 3, both models involved a high probability to look at the red ROI centered at the bridge of the nose between the two eyes with a large attention window at their first fixation (Priors), which could be the source of the association between looking at the eye region and HP. However, while the eyes-focused model always started with a fixation at this red ROI, the nose-focused model had a 47% probability to start with a fixation at a similar location but with a small attention window (Blue ROI). This could be related to why nose-focused patterns sometimes could be related to decreased HP. Similarly, after the red ROI, the eyes-focused model had high probabilities to switch between the two individual eyes with a small attention window, which could explain why a very eyes-focused pattern may also sometimes be associated with reduced HP. DNN+HMM allowed us to model the association between fixation locations and attention window sizes, which could not be readily observable in human eye movement data. This information helped us better understand the relationship between face scanning pattern and HP. Also, this relationship was not observed at the early featural processing stage of the model (LastConv layer), suggesting that the information use associated with eye fixations may be better represented at intermediate and late processing stages.

The XAI visualization further supported our speculation. At the early perceptual stage, models in both eyes- and nose-focused groups used information at the middle SF scale from the eye region and information at the low SF scale from the face center. This suggested that the two groups had similar information use from similar regions of a face regardless of the face scanning pattern differences, consistent with some previous studies (Miellet et al., 2013). At the intermediate stage, some active nodes in the FC layer encoded only local features, while some others encoded information across the whole face. This was observed in both groups, suggesting again that both groups have developed local and holistic internal representations during training, and the difference between the two groups was in the temporal dynamics of how these representations were used for face recognition as summarized in their respective HMMs (Figure 3).

In conclusion, by applying DNN+HMM to modeling face recognition, we have provided a computational account for the relationship among HP, face scanning pattern, and face recognition performance. More specifically, the model accounted for the positive associations between HP and eyes-focused face scanning pattern/face recognition performance observed in the literature regardless of whether the standard or the complete composite design was used to measure HP. This finding suggested that the inconsistent results obtained in the literature from human data may be related to insufficient testing power. Interestingly, we also observed a quadratic relationship between HP and face scanning pattern, where models being highly eyes-focused or highly nose-focused had lower HP. By inspecting the relationships between fixation locations and attention window size in the model and the model visualization through XAI methods, we found that the eyes- and nose-focused models have both developed local and holistic internal representations during training, and their difference was in the temporal dynamics of how these representations were used as summarized in their respective HMMs. Our study thus demonstrated well how computational modeling could help unravel the information processing mechanisms underlying cognition not readily observable in human data.

Acknowledgments

We are grateful to Research Grant Council of Hong Kong (CRF #C7129-20G and GRF #17608621). This work was also supported by a Strategic Research Grant from City University of Hong Kong (Project No. 7005995).

References

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

Cheng, Z., Hayward, W. G., Chan, A. B., & Hsiao, J. H. (2018). Optimal face recognition performance involves a balance between global and local information processing: Evidence from cultural difference. In *40th Annual Meeting of the Cognitive Science Society: Changing Minds, CogSci 2018 (*pp. 1476-1481*)*. The Cognitive Science Society.

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

Chung, H. K., Leung, J. C., Wong, V. M., & Hsiao, J. H. (2018). When is the right hemisphere holistic and when is it not? The case of Chinese character recognition. *Cognition, 178*, 50-56.

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

Davis, J., McKone, E., Zirnsak, M., Moore, T., O'Kearney, R., Apthorp, D., & Palermo, R. (2017). Social and attention-to-detail subclusters of autistic traits differentially predict looking at eyes and face identity recognition ability. *British Journal of Psychology*, *108*(1), 191-219.

DeGutis, J., Wilmer, J., Mercado, R. J., & Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition*, *126*(1), 87-100.

Gao, Z., Flevaris, A. V., Robertson, L. C., & Bentin, S. (2011). Priming global and local processing of composite faces: revisiting the processing-bias effect on face perception. *Attention, Perception, & Psychophysics*, *73*, 1477-1486.

Gauthier, I., & Bukach, C. (2007). Should we reject the expertise hypothesis?. *Cognition, 103*(2), 322-330.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261-2271.

Hole, G. J. (1994). Configurations! Factors in the perception of unfamiliar faces, *Perception*, *23* (1), 65–74.

Hsiao, J. H., An, J., Hui, V. K. S., Zheng, Y., & Chan, A. B. (2022). Understanding the role of eye movement consistency in face recognition and autism through integrating deep neural networks and hidden Markov models. *npj Science of Learning*, *7*(1), 28–28.

Hsiao, J. H., An, J., Zheng, Y., & Chan, A. B. (2021a). Do portrait artists have enhanced face processing abilities? Evidence from hidden Markov modeling of eye movements. *Cognition*, *211*, 104616.

Hsiao, J. H., & Galmar, B. (2016). Holistic processing as measured in the composite task does not always go with right hemisphere processing in face perception. *Neurocomputing*, *182*, 165-177.

Hsiao, J. H., Lan, H., Zheng, Y., & Chan, A. B. (2021b). Eye movement analysis with hidden Markov models (EMHMM) with co-clustering. *Behavior Research Methods*, *53*(6), 2473–2486.

Miellet, S., Caldara, R., & Schyns, P. G. (2011). Local Jekyll and global Hyde: The dual identity of face identification. *Psychological Science*, *22*(12), 1518-1526.

Miellet, S., Vizioli, L., He, L., Zhou, X., & Caldara, R. (2013). Mapping face recognition information use across cultures. *Frontiers in Psychology*, *4*, 34.

Omigbodun, A., & Cottrell, G. (2013). Is Facial Expression Processing Holistic?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, No. 35).

Rezlescu, C., Susilo, T., Wilmer, J. B., & Caramazza, A. (2017). The inversion, part-whole, and composite effects reflect distinct perceptual mechanisms with varied relationships to face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(12), 1961.

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological Bulletin*, *140*(5), 1281.

Richler, J. J., Gauthier, I., Wenger, M. J., & Palmeri, T. J. (2008). Holistic processing of faces: perceptual and decisional components. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(2), 328.

Richler, J. J., Wong, Y. K., & Gauthier, I. (2011). Perceptual expertise as a shift from strategic interference to automatic holistic processing. *Current Directions in Psychological Science*, *20*(2), 129-134.

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition*, *21*(2), 139-253.

Rossion, B., Gao, Z., Flevaris, A., Robertson, L., & Bentin, S. (2013). Global processing of Navon stimuli primes the general (face) congruency effect but not the standard composite face effect. *Journal of Vision*, *13*(9), 98-98.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the *IEEE International Conference on Computer Vision* (pp. 618-626).

Wolf, L., Hassner, T., & Taigman, Y. (2010). Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(10), 1978-1990.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747-59.

Zheng, Y., & Hsiao, J. H. (2023). Differential audiovisual information processing in emotion recognition: An eye-tracking study. *Emotion*, *23*(4), 1028.

Zheng, Y., Que, Y., Hu, X., & Hsiao, J. H. (2022). Predicting reading performance based on eye movement analysis with hidden Markov models. In *Proceedings of the 2022 International Conference on Advanced Learning Technologies (ICALT)* (pp. 172-176). IEEE.

Zhong, N., Hsiao, J. H., Zhou, G., & Hayward, W. (2024). Association of idiosyncratic eye-movement patterns with holistic processing of faces as measured by the composite face effect and the face inversion effect. *Visual Cognition*.

Zhou, G., Cheng, Z., Zhang, X., & Wong, A. C. N. (2012). Smaller holistic processing of faces associated with face drawing experience. *Psychonomic Bulletin & Review*, *19*, 157-162.