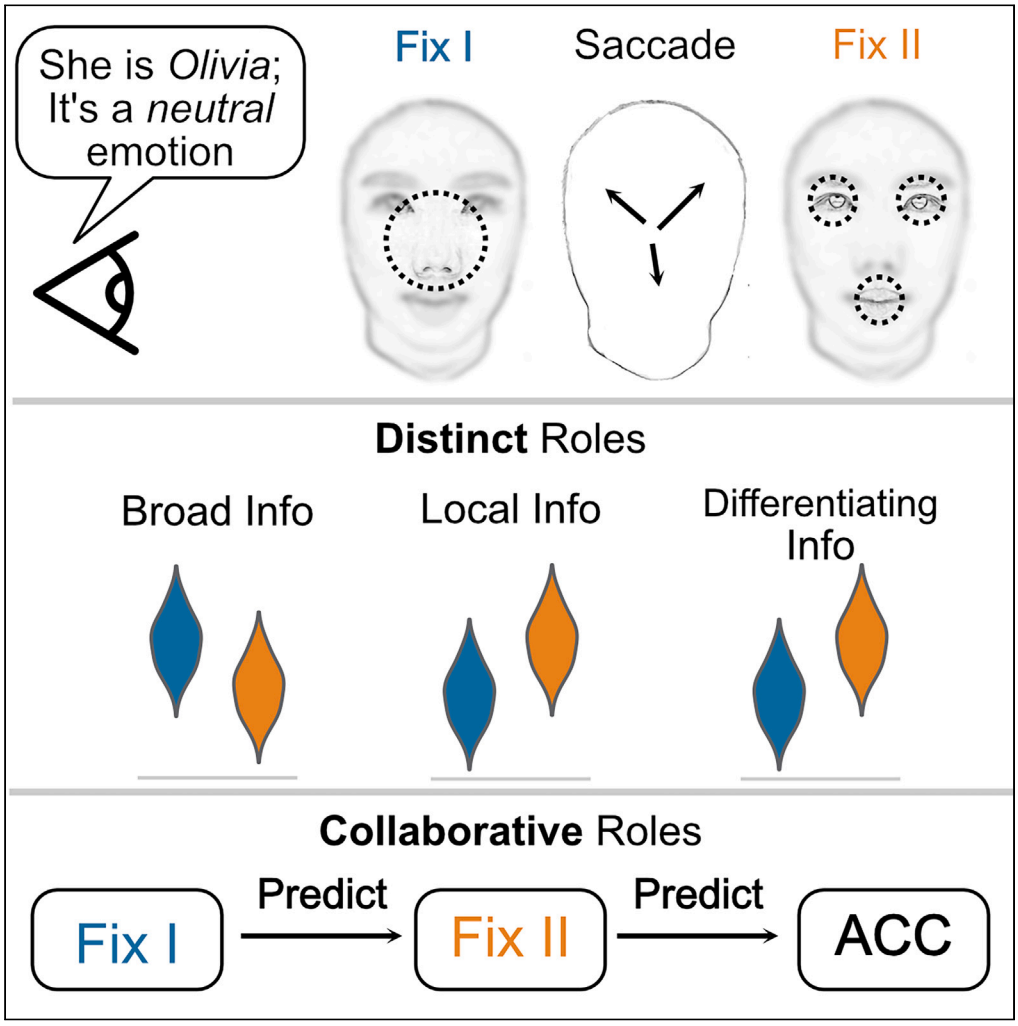
iScience

CellPress
OPEN ACCESS

Article

Specified functions of the first two fixations in face recognition: Sampling the general-to-specific facial information



Meng Liu, Jiayu Zhan, Lihui Wang

lihui.wang@sjtu.edu.cn

Highlights

The first two fixations for face perception have a central-to-divergent pattern

Fix I samples the broad facial information and Fix II samples the local information

Fix II correlates more with the differentiating information for discriminating faces

The general-to-specific information is obtained by Fix I and Fix II's collaboration

Liu et al., iScience 27, 110686
September 20, 2024
https://doi.org/10.1016/j.isci.2024.110686

iScience

CellPress
OPEN ACCESS

Article

Specified functions of the first two fixations in face recognition: Sampling the general-to-specific facial information

Meng Liu,[1,2,3] Jiayu Zhan,[4,5,6] and Lihui Wang[1,2,3,7,*]

SUMMARY

Visual perception is enacted and constrained by the constantly moving eyes. Although it is well known that the first two fixations are crucial for face recognition, the function of each fixation remains unspecified. Here we demonstrate a central-to-divergent pattern of the two fixations and specify their functions: Fix I clustered along the nose bridge to cover the broad facial information; Fix II diverged to eyes, nostrils, and lips to get the local information. Fix II correlated more than Fix I with the differentiating information between faces and contributed more to recognition responses. While face categories can be significantly discriminated by Fix II's but not Fix I's patterns alone, the combined patterns of the two yield better discrimination. Our results suggest a functional division and collaboration of the two fixations in sampling the general-to-specific facial information and add to understanding visual perception as an active process undertaken by structural motor programs.

INTRODUCTION

When we see a face, we quickly get various information about this person such as identity and emotional state. To achieve quick and accurate recognition the task-relevant information on the face has to be efficiently sampled and analyzed.[1–3] This is coupled with eye movements, which direct the gaze center toward a specific region so that the task-relevant information can reach the fovea and gain optimal processing.[4,5] Even when free eye movements are constrained, there are microsaccades within the fixation which have a similar information sampling function as the saccades during free-viewing by bringing the task-relevant information close to the fixation locus[6,7]

A large body of eye-movement studies has investigated the spatial distribution of the fixations on the face image, revealing a T-shaped concentration covering the eye-nose-mouth area[4,8,9] (also see[10] for individual differences within this area). Although these studies have provided important insights into "what" information on the face image is sampled to support recognition, it remains unsettled "how" the information is sampled and interpreted. The answer to the "how" question is often explained as neural computations or representations of the "what" information. Studies from this perspective have provided important insights into the relationship between the physical properties of the face image and the spatial-temporal neural dynamics in the brain (see a recent review[11]). However, this mapping process does not consider a fundamental body constraint: the eyes are never still and the images on the retina are always changing.[12] The information scattered on a face has to be transformed into the stimulations on the retina by the sequentially occurring eye movements.[13] Notably, a growing body of recent studies has suggested that the eye movements themselves contribute to the visual activities in the brain, which cannot be reduced to the physical properties of the images.[14–16]

Even as one of the most common visual tasks, face recognition cannot be sufficiently achieved within one single fixation. For instance, by manipulating the maximal number of fixations allowed during face viewing, Hsiao and Cottrell[17] found that the recognition performance improved significantly when the number of allowed fixations was increased from one to two, with no further improvement after more fixations had been allowed. Of note, when the numbers of fixations are strictly restricted, we speculate that a better cognitive strategy would be to direct the limited fixations to different locations rather than the same location to maximize the obtained information. In other words, from an economic perspective, why should we have the two essential fixations doing the same job? Moreover, there are individual differences in landing the first fixation toward a specific location on the face, and the recognition performance was impaired when observers were forced to maintain the first fixation away from the preferred location.[18,19] These results further suggested that the order of the fixations also matters for face recognition, because even if the preferred first location could be visited by the following fixations, forcing the first fixation away from

[1]Institute of Psychology and Behavioral Science, Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200030, China
[2]School of Psychology, Shanghai Jiao Tong University, Shanghai 200030, China
[3]Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200030, China
[4]School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, China
[5]Institute for Artificial Intelligence, Peking University, Beijing 100871, China
[6]State Key Laboratory of General Artificial Intelligence (BIGAI), Beijing 100871, China
[7]Lead contact
*Correspondence: lihui.wang@sjtu.edu.cn
https://doi.org/10.1016/j.isci.2024.110686

the preferred location would change the preferred order of the fixation sequence. Despite the functional significance of the sequentiality, the first two fixations were often treated homogenously and pooled together with other fixations to show the overall distribution on the face. No study yet to our knowledge has specified the function of each fixation and how the two fixations collaborate to achieve recognition.

Neuroimaging studies have indicated the sequential process of face recognition as information accumulation in the brain. This process can be parsed into two stages: an early stage where the initial broad information is acquired to have an overview of the face configuration, and a late stage where the detailed information is analyzed to achieve specific task goals such as recognizing the identity.[20,21] In agreement with a temporal profile, a recent model suggested that the sequence of eye movements is carried out to build up the initial perceptual hypothesis about an encountered face constrained by the task and collect information to confirm this hypothesis.[22] It is thus to be tested if the two stages of face processing can be reflected in the first two fixations. Testing the correspondence between the temporally obtained information and the first two fixations will not only show the specific functions of the first two fixations for face recognition but also advance the current understanding by bridging the "what" and "how" aspects of the information sampling. Moreover, the functions of the first two fixations in implementing the information accumulation can provide grounding for the neural representations underlying face recognition.

Here we investigated the specific functions of the first two fixations for face recognition. To show if the functions can be generalized to different recognition tasks, fixation patterns were examined in two task contexts: an identity recognition task and an emotion recognition task. Using a data-driven method, we first delineated the geographical distribution of the initial two fixations, with the first fixation (Fix I) mainly clustered along the nose bridge while the second fixation (Fix II) was more widely dispersed over key areas such as the eyes, nostrils, and lips. Building upon previous studies, we hypothesized that the distinct spatial distributions of the first two fixations reflect an optimized information sampling process for face recognition, with Fix I more involved in harvesting general facial information to build up an initial hypothesis about this face (e.g., face configuration related to an identity or emotion category), whereas Fix II more involved in processing fine details requisite for the specific recognition task (e.g., distinguishing John vs. Mike or happy vs. sad) to narrow down and confirm the perceptual hypothesis. We evaluated this research hypothesis by analyzing the correspondence between the fixation patterns and the information on the face images. Firstly, we calculated the distance between the fixation location and the regions of interest (ROIs) of the face images (eyes, nose, and mouth). We found that Fix I tended to land at a location that allows the entire face to be covered by the visual span, whereas Fix II tended to land at a location where critical local information can be scrutinized. Secondly, we quantified the information that differentiated a specific face image from other images. We found that Fix II was more related to the differentiating information than Fix I, and contributed more to the behavioral recognition performance. Thirdly, we assessed the collaboration between the two fixations. We found that the combined patterns of Fix I and Fix II were better in predicting the face categories than the patterns of Fix I or Fix II alone.

RESULTS

Twenty-eight participants completed an identity task and an emotion task (Figure 1) in a counterbalanced order, with their free eye movements recorded during the task. A picture of a target face was presented at the screen center for 1.5s (duration chosen according to Wang et al.,[15] after which a task frame was presented requiring a response from the participant. In the identity task, participants were asked to recognize the identity of the target picture by choosing the right answer from 7 alternative pictures. In the emotion task, they were asked to recognize the emotion by choosing the right answer from 7 alternative labels ("anger," "disgust," "fear," "happiness," "neutral," "sad," and "surprise"). For our research purpose, data analysis was focused on the first two fixations during the presentation of the target face. To avoid biasing the first fixation to the screen center, a fixation dot was randomly presented at one of the screen corners before the target (Figure 1). Participants exhibited high accuracies in both tasks (identity task: mean = 97.85%, SE = 0.59%; emotion task: mean = 83.70%, SE = 1.10%). The onsets of the first two saccades: mean = 217.3 ms, SE = 1.2 ms and mean = 502.0 ms, SE = 2.8 ms; the lag between the two: mean = 284.7, SE = 2.4 ms.

Distinct patterns of the first two fixations revealed by data-driven clustering

We first measured the preferred landing locations of the first two fixations (termed Fix I and Fix II thereafter) on the face image. To this end, we applied an unsupervised machine learning technique, K-means clustering, to the fixation locations pooled over trials and participants. We found that, for Fix I, participants consistently preferred only one landing location across trials (i.e., optimal K = 1, decided by the silhouette score, see Rousseeuw[23] in quantifying the goodness of fit for the K-means model; Figure 2A); for Fix II, participants preferred three landing locations across trials (i.e., optimal K = 3, Figure 2A). The specific locations of the clustered fixations from one example participant are illustrated in Figure 2B.

We summarized the preferred landing locations of all participants in Figure 2D. The centroids of the identified clusters are illustrated on an example face (Figure 2D left). Then, we transformed the K-means scatters into density layers using kernel density estimate (KDE) and overlaid the density layers of Fix I (Figure 2D middle) and Fix II (Figure 2D right) onto an example face image. The graphical results indicated that the clusters of Fix I were mainly located along the line from the middle of the eyebrows to the nose tip, and the clusters of Fix II were mainly located in the critical local regions such as the eyes, nostrils, and lips.

The above independent patterns of the first two fixations showed overlapped areas, which might have been taken as that Fix I and Fix II landed at the same location. However, it should be noted that the overlapped areas were not necessarily from the same trial. To verify the distinct patterns, we further visualized the within-trial transition pattern from Fix I and Fix II. We color-coded the saccade trajectory from Fix I and Fix II at the single-trial level and mapped the trajectory onto the face image (see Figure 2E for data from one participant and Figure S1 for

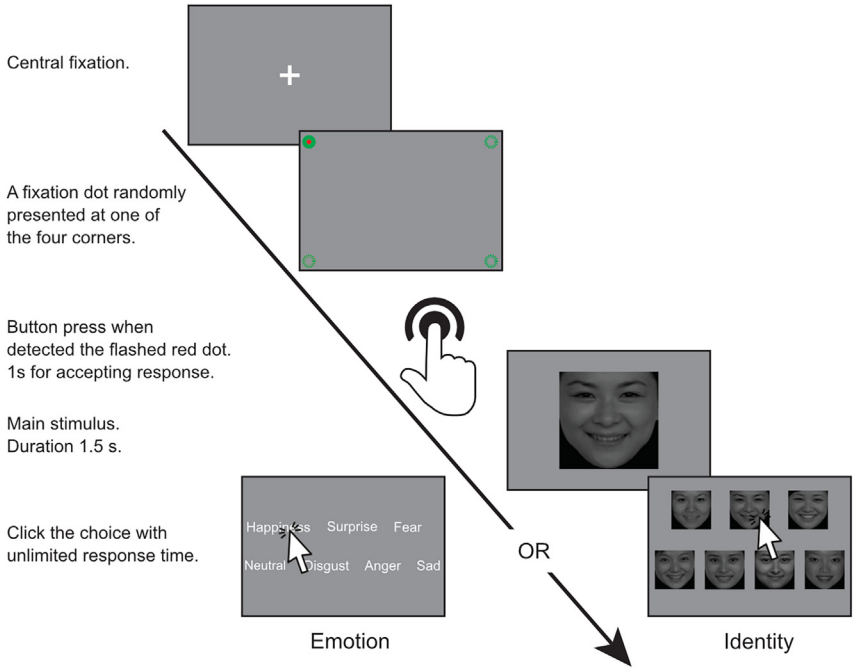


Figure 1. Experimental design and stimuli sequence

Each trial began with a central fixation cross, followed by a green fixation dot at a random location of the four corners. Note that all four possible positions are marked by the green circles here but only one green dot was presented in a specific trial. In 10% of all trials, a red dot inside the green dot was presented, and participants were asked to detect it by a button press. Then a target face was presented, during which participants were asked to view with free eye movements. In the identity task, participants were asked to choose the correct image for the target face. In the emotion task, participants were asked to choose the correct emotion label for the target face (labels shown here are in English but were in Chinese in the experiment). Responses were given by mouse click without a time limit.

data from other participants). The geographical pattern suggested a central-to-divergent transition from Fix I to Fix II on the image (see the section below and Figure 3C for statistical evidence).

The distinct roles of the first two fixations in sampling general-to-specific facial information

The clustering of the fixation distribution suggested a centrality of Fix I on the face, whereas Fix II diverged into the key areas such as the eyes and mouth. Building upon previous studies on the critical physical properties of the face images for recognition,[25,26] the distinct patterns between Fix I and Fix II may function to sample the general-to-specific information for face perception. We assessed the "general vs. specific" information in correspondence to the functional division of Fix I and Fix II on the following aspects: 1) how the landing locations of the two fixations were biased for the broad versus local facial information contained within a face; 2) how the patterns of fixations relate to the information differentiating the current face from other faces; 3) how the patterns of fixations contribute to the performance of behavioral discrimination.

Fix I covers broad facial information, Fix II targets local facial information

In a first step, we asked if and how the landing locations of Fix I and Fix II were involved in sampling the broad information spatially covering all key ROIs (eyes, nose, and mouth) or were biased toward the local details of a specific ROI. Specifically, landing at a central-middle location accessible to all ROIs would be beneficial for having the whole face within the visual span whereas the detailed information of a specific ROI would be nevertheless compromised because the visual resolution decreased from the fovea to the peripheral visual field. By contrast, landing at one of the ROIs would be beneficial for getting fine details of this ROI whereas the broad information such as the relational information between this ROI and other ROIs would be nevertheless compromised.

We predicted that Fix I would land at a location accessible to all key ROIs, whereas Fix II would land at one of the ROIs. We selected four key ROIs (left eye, right eye, nose, and mouth) for each face image based on facial landmarks detected with a machine learning tool – Dlib[24]; Figure 3A, left). We calculated the distance from a specific location to each of the four ROIs (namely four distances, in degrees of visual angle), and defined two types of distance: the shortest distance among the four distances (i.e., Shortest-Dis) and the sum-up of the four distances (i.e., Sum-Dis, Figure 3A, right). Specifically, Shortest-Dis quantified to which extent the fixation location was closer to one of the four ROIs than

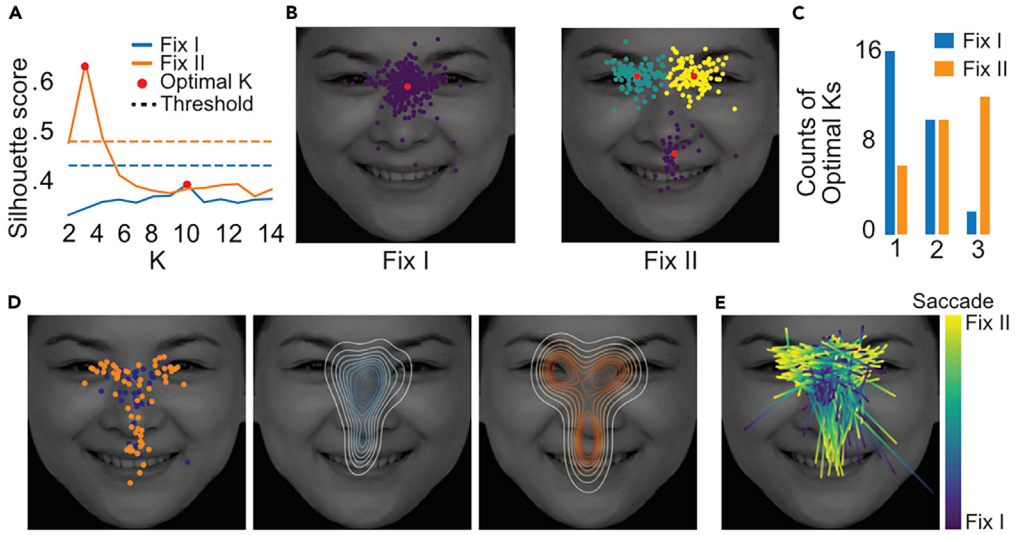


Figure 2. Clustering results of the first two fixations

(A) The goodness of fit (i.e., silhouette score,[23] is shown as a function of the prior number of clusters (K) for Fix I (blue line) and Fix II (orange line) for one example participant. The dashed lines denote the threshold for deciding the optimal K value. The red dots denote the optimal Ks, which achieved the maximum silhouette scores.

(B) The clusters of Fix I (left) and Fix II (right) are obtained from the optimal K-means model for the same example participant. Purple, cyan, and yellow dots indicate the individual fixations pooled over trials; red dots indicate the geometrical centers of the clusters.

(C) The counts of optimal Ks of all participants for Fix I (blue bars) and Fix II (orange bars).

(D) The distribution of the cluster centroids for the first two fixations (left panel, each dot represents the centroid of the estimated cluster of fixations across trials for a specific participant, i.e., the red dots in B, blue for Fix I, orange for Fix II). The separate patterns of Fix I (middle panel) and Fix II (right panel) are also shown in the form of density layers.

(E) The saccade trajectory from Fix I to Fix II at the single-trial level. The color-coded lines represent single-trial fixation pairs from one example participant, where the transition from blue to yellow represents the trajectory from Fix I to Fix II.

the other three, and Sum-Dis quantified the cost of the fixation to get access to all four ROIs. Our analysis showed that the image pixels with the lowest Sum-Dis were localized at the center of the face image, whereas the pixels with the lowest Shortest-Dis were localized at the four ROIs (Figure 3A, right). The centrality versus divergence patterns here echoed the distinct clustering patterns between Fix I and Fix II illustrated in Figure 2.

Further statistical analysis revealed that Shortest-Dis of Fix I was higher than Shortest-Dis of Fix II, $t(27) = 6.15$, $p < 0.001$, Cohen's $d = 1.16$ (Figure 3B, left). By contrast, Sum-Dis was lower for Fix I than for Fix II, $t(27) = 2.99$, $p = 0.006$, Cohen's $d = 0.57$ (Figure 3B, right). We also replicated the results in each of the two tasks. Specifically, Shortest-Dis of Fix I was higher than Shortest-Dis of Fix II in both the identity task, $t(27) = 6.33$, $p < 0.001$, Cohen's $d = 1.20$, and the emotion task, $t(27) = 5.04$, $p < 0.001$, Cohen's $d = 0.95$; Sum-Dis was lower for Fix I than for Fix I in both the identity task, $t(27) = 3.11$, $p = 0.004$, Cohen's $d = 0.59$, and the emotion task, $t(27) = 2.32$, $p = 0.028$, Cohen's $d = 0.44$. Note that the small effect size in the emotion task according to the prior criteria ($d < 0.48$, see STAR Methods) asked for further examinations in future studies. These results suggested that, regardless of task type, the location of Fix I was more beneficial for sampling the broad information over all ROIs, whereas the location of Fix II was more beneficial for sampling the local information of one ROI.

Then we verified if Fix II was more distant from the face center than Fix I at the single-trial level, as shown by the graphical results of the central-to-divergent tendency (cf., Figure 2E). To test this prediction, we chose the location on the face image that had the shortest Sum-Dis as the face center (Figure 3A). For each trial, we calculated two distances – the distance from the face center to Fix I and to Fix II. Across trials, we compared these two single-trial distances with paired *t* test, independently for each participant. Given the individual-level statistical analysis, the Bayesian population prevalence (BPP,[27] was introduced to assess the statistical reliability. Statistical analysis collapsed over the two tasks showed that Fix II was more distant from the face center than Fix I ($p < 0.05$ for 23 out of 28 participants, see Figure 3C left), with the maximum a posteriori probability (MAP) estimate of the population prevalence of 23/28 achieved 0.84, 95% highest posterior density interval, HPDI = [0.63,0.93]. That is, given the data, the probability that the population prevalence was greater than 63% was higher than 95%. Separate statistical analysis showed $p < 0.05$ for 20 out of 28 participants in the identity task, and $p < 0.05$ for 21 out of 28 participants in the emotion task.

Fix II correlates more than Fix I with differentiating facial information

As the results of distance revealed the general versus the specific information processing of a face, a further prediction is that the specific information can differentiate one specific face from other faces, and such differentiating information can be used to achieve the behavioral

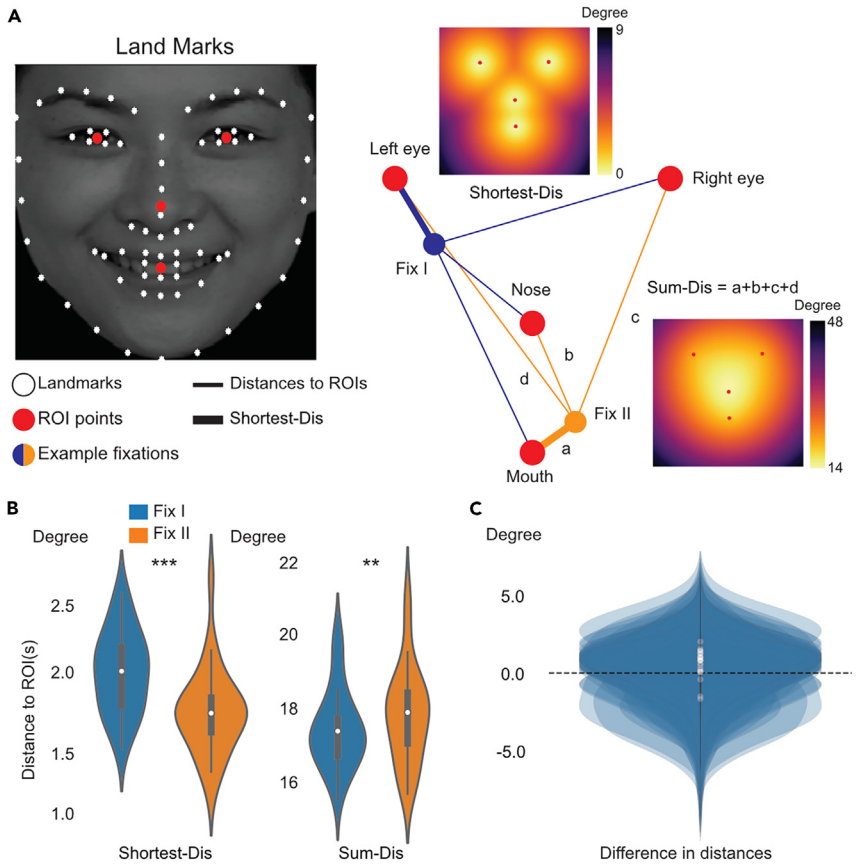


Figure 3. Definition of face ROIs and analysis of distance

(A) Left: the facial landmarks (marked by white dots) detected with the machine learning model – Dlib,[24] were used to identify the ROIs (eyes, nose, mouth). The red dots denote the geometric center of each ROI. Right: the red disks indicate the centroids of the four ROIs, the blue and orange disks are shown as an example of Fix I and Fix II, respectively. The thin lines (blue for Fix I and orange for Fix II) indicate the distances from a fixation location to an ROI, and thick lines indicate the Shortest-Dis. The heatmaps illustrate to which extent each image pixel has a low Shortest-Dis (the upper heatmap) and a low Sum-Dis (the lower heatmap).

(B) The two types of distance for the first two fixations (left panel: Shortest-Dis, right panel: Sum-Dis; blue for Fix I, orange for Fix II). The violin plots represent the distributions of the participant-level distances (**$p < 0.01$, ***$p < 0.001$, paired t tests).

(C) Difference in distance from Fix II and Fix I to the face center. Each violin plot represents the distribution of trial-level measurement for a specific participant.

discrimination of identity or emotion among the alternatives. In a second step, we defined the Differentiating Features by quantifying to which extent one face was different from other faces and investigated how the patterns of fixations correlated with the Differentiating Features. Based on our hypothesis, we predicted that the patterns of Fix II would be more related to the Differentiating Features than Fix I. Specifically, for each target face, we calculated the pair-wised difference of pixel values between the target face and each of the other faces. By averaging the pair-wised differences, we obtained a feature map for each target face, quantifying to which extent the current target face image differed from other face images (Figure 4A).

An example of the differentiating features for a face and the feature-fixation overlay for this example is illustrated in Figure 4A (the two columns on the right), where the fixations corresponded to the pixels with high differentiating values (in bright color). We then investigated how the differentiating features corresponded to the patterns of Fix I and Fix II. For each fixation group (Fix I vs. Fix II), the correspondence was measured by calculating the cross-correlation between the vectorized values of the feature map and the vectorized values of the fixation density map. Statistical analysis of the cross-correlation coefficients showed that the density map of Fix II correlated more with the differentiating features than Fix I (0.20 vs. 0.24), $t(27) = 6.73$, $p < 0.001$, Cohen's $d = 0.97$ (Figure 4B, left panel).

Mirroring the distance findings, the correlation patterns between the density map and differentiating features are consistent across the two tasks. Fix II correlated more with the differentiating features than Fix I in both the identity task (0.55 vs. 0.70), $t(27) = 6.37$, $p < 0.001$, Cohen's $d = 1.16$ (Figure 4B, lower middle panel), and the emotion task (0.55 vs. 0.67), $t(27) = 4.76$, $p < 0.001$, Cohen's $d = 0.81$ (Figure 4B, middle and right panels).

The above analysis of distance and feature-fixation correlation focused on the first two fixations. As a comparison, we also analyzed the following 3-5th fixation, which showed a similar pattern to the second fixation (see Figure S2 and Figure S3).

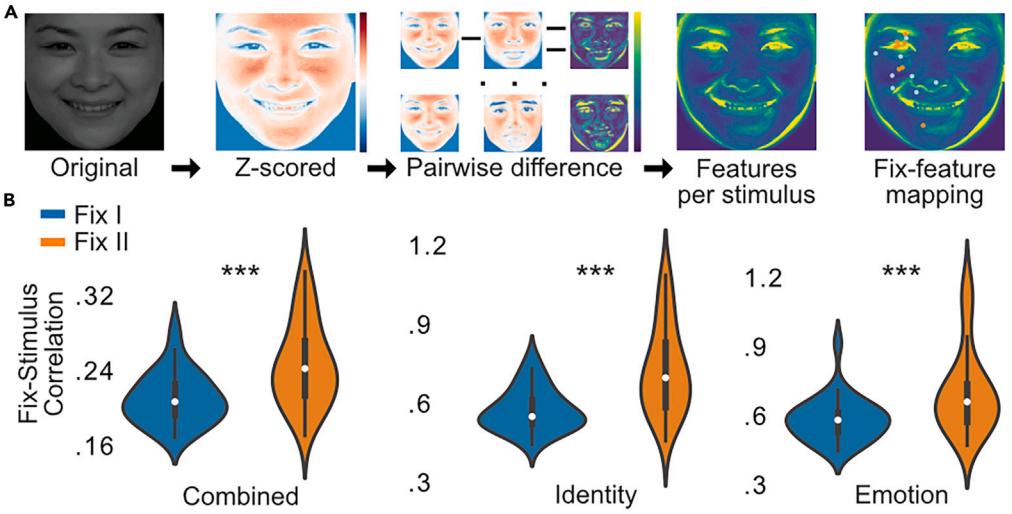


Figure 4. The facial feature maps and the fixation-feature correlations

(A) The flow charts show the procedures of calculating the Differentiating Features for each target face image. The second column from the right illustrates a map of the Differentiating Features for example face (the first column from the left, high values in bright color). The first column from the right illustrates the overlay of Fix I (blue dots) and Fix II (orange dots) on the map of the Differentiating Features.

(B) The cross-correlation (Z-transformed) between the fixation density map and the Differentiating Feature map (blue for Fix I, orange for Fix II). The violin plots illustrate the distribution of the participant-level correlation coefficients (***$p < 0.001$, paired *t-tests*).

Behavioral discrimination is better predicted by Fix II

In a third step, we asked if and how the two fixations contributed to the behavioral performances of face recognition. We addressed this by investigating which one of the two fixations' patterns was more distinctive between the correct and incorrect recognition responses. We focused on the emotion task, as the accuracies of the identity task approached the ceiling (mean accuracy = 97.85%, Figure 5A), leaving insufficient trials with an incorrect response for analysis. Specifically, we calculated the correlation between the Differentiating Feature map and the fixation density map with the same procedure as above, but the calculation was performed for each of the 7 emotion categories, and trials with correct and incorrect responses respectively.

The paired *t*-test comparing the correspondence between the correct and the incorrect trials showed a significant difference for Fix II, $t(27) = 2.34$, $p = 0.026$, Cohen's $d = 0.44$, but not for Fix I, $t(27) = 0.51$, $p = 0.616$ (Figure 5B). Note that the small effect size ($d < 0.48$) asked for further examinations in future studies. That is, relative to trials with incorrect recognitions, trials with correct recognitions had higher correlations between the Differentiating Features and the fixation patterns, and this was observed only for Fix II but not Fix I. To test the reliability of the results, we further assessed the difference for the following 3-5th fixations, none of which showed any significant effects (3rd: $p = 0.776$, 4th: $p = 0.230$, 5th: $p = 0.952$). These results suggested that the successful behavioral discrimination of the face categories (emotion category here) was dependent on how well Fix II but not Fix I corresponded to the Differentiating Features. In an extension of the above distance and the feature-fixation correspondence, these results suggested that the specific information sampled by Fix II can be used to achieve behavioral discrimination.

The collaboration between Fix I and Fix II

So far, by treating Fix I and Fix II independently, we have shown that Fix I and Fix II had distinct roles in sampling the facial information to support behavioral recognition. A remaining question is: were they collaborative in collecting the task-guided information (i.e., identity information in the identity task and emotion information in the emotion task)? The lower correlation with the differentiating features and the lack of power in predicting successful recognition response might lead to an underestimated task relevance of Fix I. In other words, the general facial information collected by Fix I may not have contained any task-relevant information. Alternatively, Fix I can also carry categorical information to form a task-guided perceptual hypothesis about the face. The information collected by Fix I itself may not be sufficiently distinguishable, but can add to the specific information collected by Fix II, and the combined information would be better in distinguishing the face categories than the specific information collected by Fix II. We thus predicted that the combined patterns of Fix I and Fix II would be more distinctive between face categories than the pattern of Fix II alone. Specifically, the classifier trained with combined patterns of the first two fixations would be better in classifying the face categories than the classifier trained with Fix II or Fix I alone. We tested this prediction by calculating the accuracy of the fixation patterns in predicting the label of the emotion categories (Figure 6A). We built three independent predictive models with different predictors respectively: the location coordinates of Fix I only (model 1), the location coordinates of Fix II only (model 2), and the location coordinates of both fixation (model 3). We adopted a random forest model to train and cross-validate the predictive

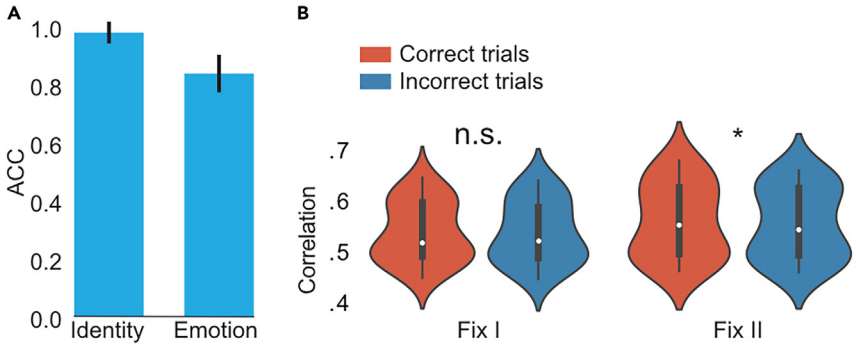


Figure 5. The contribution of Fix II to recognition performance

(A) The mean and standard deviation of the accuracies (ACC) in the two tasks.

(B) The correspondence (Z-transformed correlations) between the differentiating features and the first two fixations for correct and incorrect trials. The violin plots illustrate the distributions of the individual correlations (n.s., *$p < 0.05$, paired *t-tests*).

model. While model 1 did not show above-chance prediction accuracy, $t < 1$, the prediction accuracies of model 2 and model 3 were both above-chance: $t(27) = 3.46$, $p = 0.002$, Cohen's $d = 0.66$ for model 2 and $t(27) = 5.04$, $p < 0.001$, Cohen's $d = 0.95$ for model 3 (Figure 6B). Importantly, the prediction accuracy of model 3 was significantly higher than the prediction accuracy of model 2, $t(27) = 2.16$, $p = 0.039$, Cohen's $d = 0.41$. Note that the small effect size concerning the higher accuracy of model 3 than model 2 ($d < 0.48$) asked for further examinations in future studies. We also trained and cross-validated other models with the fixation data and found the same patterns of results: the prediction accuracy increased from model 1 to model 2, and to model 3 (Figure S4).

Here we focused the classification on the emotion labels not on the identity labels because we included 70 target identities to avoid the confounding effect of familiar faces.[28,29] The classification of a large number of identities based on the current sample size would run the risk of overfitting. As a sanity check, see Figure S4 for the results of the identity-based classification.

DISCUSSION

It is long established that face perception involves strategic fixations on the face image, gathering the visual information for recognition.[4,30] Recent findings highlighted that the functional significance of the fixations for face recognition extends beyond spatial patterns to temporal relations between the fixations.[22,31,32] It has been shown that the recognition performance could be related to the transition probability from one fixation location to the next on the face image.[32] A computational model of visual recognition suggested that the location of a fixation determines the location of the next fixation[22] (see the ref. Arizpe et al[8] and Ziman et al[33] for empirical evidence). According to this model, an initial hypothesis about a stimulus (e.g., face identity) is built up based on the information sampled by the preceding fixation, which leads the next fixation to the expected location to confirm the initial hypothesis. Together with these studies, the current results suggested a spatio-temporal information integration process for face recognition that can be optimized by structural fixation sequence, with the individual fixations serving distinct and collaborative functions.

Previous studies have shown a central tendency of the first fixation during face perception.[17–19,29] In Hsiao and Cottrell[17] where the face size was smaller than the face size shown here, the first fixation was located around the center of the nose. In Peterson and Eckstein[18] where the faces had a similar size as the current study, the first fixation was located below the eyes while above the nose tip across multiple recognition tasks. This centrally localized pattern was suggested as a strategy to optimize the information integration because the whole face was covered by the perceptual span and thus all relevant information can be acquired.[17,18] Consistent with these findings, here the graphical analysis of the fixation locations also showed that Fix I was clustered along the nose bridge. As a further step, here the clustering results showed the graphical pattern of the second fixation, which diverged to the eyes, nostrils, and lips. Importantly, the current findings provide a mechanistic account for the central vs. divergent distributions. The distance results suggested that the centrally localized pattern of Fix I was to cover the broad facial information of the key ROIs while the divergent pattern of Fix II was to target the specific ROIs for local information. Moreover, the feature-fixation correlation results showed that Fix II was more related to the differentiating information than Fix I and contributed more to the behavioral recognition performance. These results suggested a functional division of the first two fixations, with the first fixation more involved in processing the general configuration information, whereas the second fixation more involved in processing the fine information specific to a certain face. Furthermore, only the patterns of Fix II alone but not Fix I alone could predict the face categories, and the combined patterns of the two fixations yield better predictions than Fix II. These results suggested that the general information gathered by Fix I, although not significantly discriminative itself, adds to the differentiating information gathered by Fix II to achieve statistically significant discrimination.

The central versus divergent distribution and the general versus specific information processing between the first two fixations raised the holistic versus analytic processing in face recognition.[34] While holistic processing combines the different parts of the face into a whole, analytic

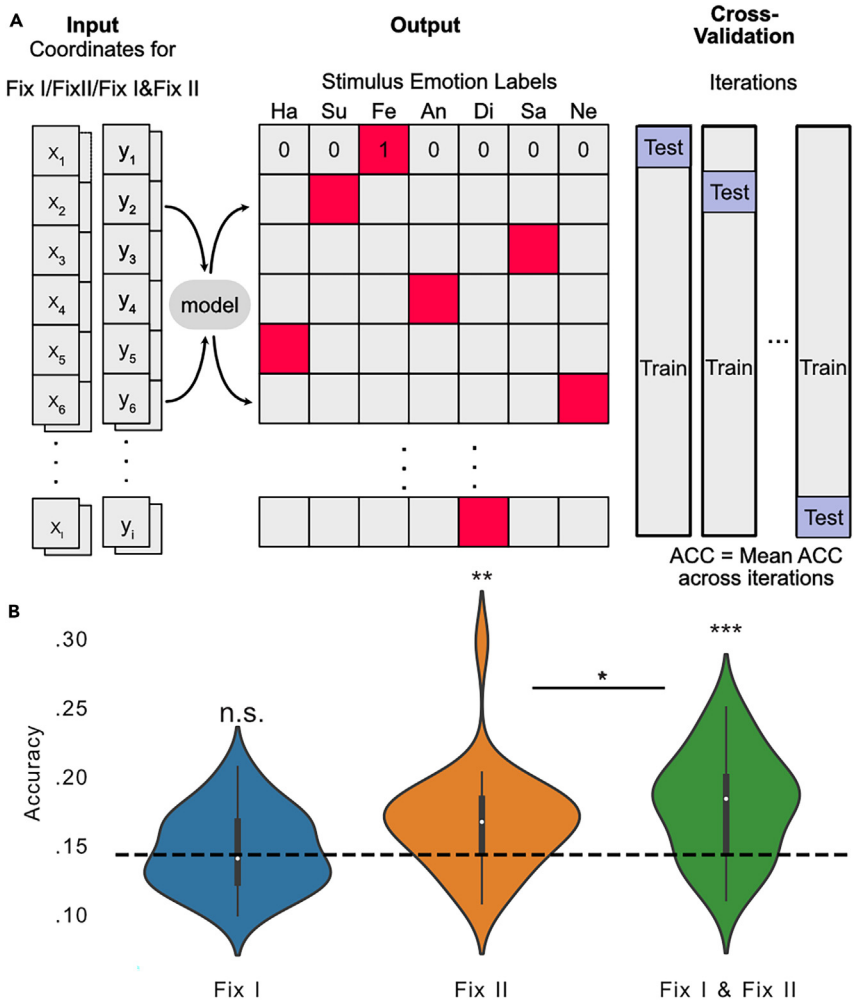


Figure 6. Predictions of the emotion categories based on the locations of Fix I, Fix II, and the two combined

(A) The input-output data for the model training and the cross-validation procedure.

(B) The cross-validation accuracy for the random forest model in cross-validation. The violin plots illustrate the distributions of the cross-validation performance of the individual participants (*$p < 0.05$, paired *t* test; n.s., **$p < 0.01$, ***$p < 0.001$, one-sample *t-tests*).

processing narrows down the visual analysis to the local information. Fixations clustered at the center of a face were taken as evidence for holistic processing, whereas fixations clustered at the local regions were taken as evidence for analytic processing.[25,31] It has been shown that analytic processing was preceded by holistic processing during face perception.[35] Moreover, face recognition performance was more related to analytic processing but less so to holistic processing.[32] Consistent with this finding, the recognition performance in the emotion task here could be better predicted by the pattern of Fix II than Fix I. Along this line, the first fixation was for holistic processing and the second fixation was for analytic processing.

Although this study's results were based on East Asian observers viewing East Asian faces, the functions of the first two fixations are consistent with previous studies on West Caucasian observers during the viewing of West Caucasian faces.[18,35] However, fixation patterns were shown to be different between cultural groups.[9,36] In a task of race recognition, West Caucasians showed a scattered pattern of fixations over the eyes and the mouth whereas East Asians tended to focus the fixations on the central region of the face.[9] In a task of emotion recognition, the fixations of West Caucasians were distributed evenly over the eyes, the nose, and the mouth whereas the fixations of East Asians were mainly distributed over the eyes and the nose.[36] It should be noted that these distinctive patterns could be due to the different weights (e.g., in terms of duration) of the individual fixations, as the individual fixations were pooled over to show the spatial pattern. The distinct functions of the individual fixations proposed here thus can be taken to tease apart the cultural differences of the fixation patterns.[37] Nevertheless, the generalizability of the current findings to different cultural groups still needs to be tested.

Time-resolved neuroimaging evidence has shown the early and late stages of face processing.[21,38] A magnetoencephalographic (MEG) study[21] showed two phases responsive to faces during the first 500 ms post-image presentation: an early phase (100–200 ms) associated

with the low-level information of the image, and a later phase (200–300 ms) associated with the high-level information of the face identity. In agreement with this sequential neural processing, the current findings suggested that the transition from general to specific processing stages was reflected in the fixation sequences. However, it should be noted that the exact onset of the two stages was subject to the context (e.g., the initial fixation started from the center in the MEG study whereas had to be redirected from the peripheral to the center in the current study). Considering that the current experiment design entailed longer fixation latencies, the lag between the two fixations (on average 284.7 ms) here intuitively fit with the sequential properties of the two MEG time ranges. Another account could be that Fix I was also involved in identity-related processing (e.g., configurational information for identity recognition), which directed Fix II into more detailed processing. The shared temporal characteristic of face perception points to a link between the fixation sequence and the brain activity revealed by recent studies where the face-related neural representations were tuned to the fixation sequences that were used to explore the face images.[14,15] Importantly, the idiosyncratic fixation patterns of the same face were coupled with idiosyncratic neural tunings among observers,[14] and the coupling between the fixation sequence and the neural representation was observed in the absence of image input.[15] Therefore, the fixation sequence and the neural response may not simply be two independent aspects of the same information sampling process but have interactively developed from the experience of visual exploration.[39,40]

The temporal course of general-to-specific information processing reflected in the fixation sequence here fits with the general principle of the "coarse-to-fine" processing in visual perception[35,41,42] beyond face perception.[43,44] From a dualistic view, the neural activities can be taken as the inner representations of the physical properties of the outer images. However, this view is challenged by alternative suggestions and growing evidence that the neural representation is grounded by action.[45,46] Building on previous work, our findings suggest fixation sequences as the grounding for neural representations of faces and shed light on a framework covering the physical image, the eye movements, and the neural representations for visual perception.

Limitations of the study

We focused on the first two fixations, yet it should not be taken that the fixations after the first two fixations did not contribute to the recognition at all. One account we speculated here is that, apart from the initial fixations, the following fixations may provide supplementary information. The supplementary information could be used to further confirm the initial hypothesis on the task-defined recognition (e.g., identity, emotion),[22] and hence might contribute more to the recognition confidence than to the correctness of the recognition. Also, we included only two basic recognition tasks, so caution should be taken in generalizing our findings to more abstract task goals such as judging competence,[47] beauty[48], or trustworthiness.[49]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Participants
- METHOD DETAILS
  - Stimuli and apparatus
  - Design and procedure
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Extraction of fixations
  - K-means clustering of the fixations
  - Analysis of distances
  - Feature-fixation correspondence
  - Prediction models of emotion categories

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110686.

ACKNOWLEDGMENTS

We thank Ms. Mengqiao Deng and Mr. Ruiming Zhu for assisting with data collection.

This work was supported by the National Natural Science Foundation of China 32271086 (LW, ML), and a Mercator Fellowship of the Deutsche Forschungsgemeinschaft 450600965 (LW).

AUTHOR CONTRIBUTIONS

Conceptualization: LW and ML; methodology: LW, ML, and JYZ; investigation: ML; visualization: ML; supervision: LW and JYZ; writing—original draft: LW and ML: writing—review and editing: LW, ML, and JYZ.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 4, 2024
Revised: June 14, 2024
Accepted: August 5, 2024
Published: August 6, 2024

REFERENCES

1. Schyns, P.G., Bonnar, L., and Gosselin, F. (2002). Show Me the Features! Understanding Recognition From the Use of Visual Information. Psychol. Sci. *13*, 402–409.
2. Smith, M.L., Cottrell, G.W., Gosselin, F., and Schyns, P.G. (2005). Transmitting and Decoding Facial Expressions. Psychol. Sci. *16*, 184–189.
3. Young, A.W., Hay, D.C., McWeeny, K.H., Flude, B.M., and Ellis, A.W. (1985). Matching Familiar and Unfamiliar Faces on Internal and External Features. Perception *14*, 737–746.
4. Yarbus, A.L. (1967). Eye Movements and Vision (Springer).
5. Henderson, J.M. (2003). Human gaze control during real-world scene perception. Trends Cogn. Sci. *7*, 498–504.
6. Shelchkova, N., Tang, C., and Poletti, M. (2019). Task-driven visual exploration at the foveal scale. Proc. Natl. Acad. Sci. *116*, 5811–5818.
7. Wang, Z., Meghanathan, R.N., Pollmann, S., and Wang, L. (2024). Common structure of saccades and microsaccades in visual perception. J. Vis. *24*, 20.
8. Arizpe, J., Kravitz, D.J., Yovel, G., and Baker, C.I. (2012). Start Position Strongly Influences Fixation Patterns during Face Processing: Difficulties with Eye Movements as a Measure of Information Use. PLoS One *7*, e31106.
9. Blais, C., Jack, R.E., Scheepers, C., Fiset, D., and Caldara, R. (2008). Culture Shapes How We Look at Faces. PLoS One *3*, e3022.
10. Mehoudar, E., Arizpe, J., Baker, C.I., and Yovel, G. (2014). Faces in the eye of the beholder: Unique and stable eye scanning patterns of individual observers. J. Vis. *14*, 6.
11. Duchaine, B., and Yovel, G. (2015). A Revised Neural Framework for Face Processing. Annu. Rev. Vis. Sci. *1*, 393–416.
12. Martinez-Conde, S., Otero-Millan, J., and Macknik, S.L. (2013). The impact of microsaccades on vision: towards a unified theory of saccadic function. Nat. Rev. Neurosci. *14*, 83–96.
13. Boi, M., Poletti, M., Victor, J.D., and Rucci, M. (2017). Consequences of the Oculomotor Cycle for the Dynamics of Perception. Curr. Biol. *27*, 1268–1277.
14. Stacchi, L., Ramon, M., Lao, J., and Caldara, R. (2019). Neural Representations of Faces Are Tuned to Eye Movements. J. Neurosci. *39*, 4113–4123.
15. Wang, L., Baumgartner, F., Kaule, F.R., Hanke, M., and Pollmann, S. (2019). Individual face- and house-related eye movement patterns distinctively activate FFA and PPA. Nat. Commun. *10*, 5532.
16. Parker, P.R.L., Martins, D.M., Leonard, E.S.P., Casey, N.M., Sharp, S.L., Abe, E.T.T., Smear, M.C., Yates, J.L., Mitchell, J.F., and Niell, C.M. (2023). A dynamic sequence of visual processing initiated by gaze shifts. Nat. Neurosci. *26*, 2192–2202.
17. Hsiao, J.H.W., and Cottrell, G. (2008). Two Fixations Suffice in Face Recognition. Psychol. Sci. *19*, 998–1006.
18. Peterson, M.F., and Eckstein, M.P. (2012). Looking just below the eyes is optimal across face recognition tasks. Proc. Natl. Acad. Sci. *109*, E3314–E3323.
19. Peterson, M.F., and Eckstein, M.P. (2013). Individual Differences in Eye Movements During Face Identification Reflect Observer-Specific Optimal Points of Fixation. Psychol. Sci. *24*, 1216–1225.
20. Kanwisher, J.L., and Alison Harris, N. (2004). Stages of Processing in Face Perception: An Meg Study. In Social Neuroscience (Psychology Press).
21. Vida, M.D., Nestor, A., Plaut, D.C., and Behrmann, M. (2017). Spatiotemporal dynamics of similarity-based neural representations of facial identity. Proc. Natl. Acad. Sci. USA *114*, 388–393.
22. Bicanski, A., and Burgess, N. (2019). A Computational Model of Visual Recognition Memory via Grid Cells. Curr. Biol. *29*, 979–990.e4.
23. Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65.
24. King, D.E. (2009). Dlib-ml: A machine learning toolkit. J. Mach. Learn. Res. *10*, 1755–1758.
25. Miellet, S., Caldara, R., and Schyns, P.G. (2011). Local Jekyll and Global Hyde: The Dual Identity of Face Identification. Psychol. Sci. *22*, 1518–1526.
26. Schwarzer, G., Huber, S., and Dümmler, T. (2005). Gaze behavior in analytical and holistic face processing. Mem. Cognit. *33*, 344–354.
27. Ince, R.A., Paton, A.T., Kay, J.W., and Schyns, P.G. (2021). Bayesian inference of population prevalence. Elife *10*, e62461.
28. Heisz, J.J., and Shore, D.I. (2008). More efficient scanning for familiar faces. J. Vis. *8*, 1–10.
29. Van Belle, G., Ramon, M., Lefèvre, P., and Rossion, B. (2010). Fixation patterns during recognition of personally familiar and unfamiliar faces. Front. Psychol. *1*, 1338. https://doi.org/10.3389/fpsyg.2010.00020.
30. Henderson, J.M., Williams, C.C., and Falk, R.J. (2005). Eye movements are functional during face learning. Mem. Cognit. *33*, 98–106.
31. Chuk, T., Chan, A.B., and Hsiao, J.H. (2014). Understanding eye movements in face recognition using hidden Markov models. J. Vis. *14*, 8.
32. Chuk, T., Chan, A.B., and Hsiao, J.H. (2017). Is having similar eye movement patterns during face learning and recognition beneficial for recognition performance? Evidence from hidden Markov modeling. Vision Res. *141*, 204–216.
33. Ziman, K., Kimmel, S.C., Farrell, K.T., and Graziano, M.S.A. (2023). Predicting the attention of others. Proc. Natl. Acad. Sci. *120*, e2307584120.
34. Maurer, D., Grand, R.L., and Mondloch, C.J. (2002). The many faces of configural processing. Trends Cogn. Sci. *6*, 255–260.
35. Peters, J.C., Goebel, R., and Goffaux, V. (2018). From coarse to fine: Interactive feature processing precedes local feature analysis in human face perception. Biol. Psychol. *138*, 1–10.
36. Jack, R.E., Blais, C., Scheepers, C., Schyns, P.G., and Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. Curr. Biol. *19*, 1543–1548.
37. Or, C.C.-F., Peterson, M.F., and Eckstein, M.P. (2015). Initial eye movements during face identification are optimal and similar across cultures. J. Vis. *15*, 12.
38. Liu, J., Harris, A., and Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. Nat. Neurosci. *5*, 910–916.
39. Floreano, D., Kato, T., Marocco, D., and Sauser, E. (2004). Coevolution of active vision and feature selection. Biol. Cybern. *90*, 218–228.
40. Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The Theory of Event Coding (TEC): A framework for perception and action planning. Behav. Brain Sci. *24*, 849–878.
41. Bar, M. (2004). Visual objects in context. Nat. Rev. Neurosci. *5*, 617–629.
42. Hochstein, S., and Ahissar, M. (2002). View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. Neuron *36*, 791–804.
43. Goffaux, V., Peters, J., Haubrechts, J., Schiltz, C., Jansma, B., and Goebel, R. (2011). From Coarse to Fine? Spatial and Temporal Dynamics of Cortical Face Processing. Cereb. Cortex *21*, 467–476.

44. Dobs, K., Isik, L., Pantazis, D., and Kanwisher, N. (2019). How face perception unfolds over time. Nat. Commun. *10*, 1258.
45. György Buzsáki, M.D. (2019). The Brain from inside Out (Oxford University Press).
46. Haxby, J.V., Gobbini, M.I., and Nastase, S.A. (2020). Naturalistic stimuli reveal a dominant role for agentic action in visual representation. Neuroimage *216*, 116561.
47. Todorov, A., Mandisodza, A.N., Goren, A., and Hall, C.C. (2005). Inferences of Competence from Faces Predict Election Outcomes. Science *308*, 1623–1626.
48. Zhan, J., Liu, M., Garrod, O.G.B., Daube, C., Ince, R.A.A., Jack, R.E., and Schyns, P.G. (2021). Modeling individual preferences reveals that face beauty is not universally perceived across cultures. Curr. Biol. *31*, 2243–2252.
49. Nightingale, S.J., and Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proc. Natl. Acad. Sci. *119*, e2120481119.
50. Acland, B.T., and Braver, T.S. (2014). Cili(v0.5.4). [Software]. https://doi.org/10.5281/zenodo.48843.
51. Giner-Sorolla, R., Montoya, A.K., Reifman, A., Carpenter, T., Lewis, N.A., Aberson, C.L., Bostyn, D.H., Conrique, B.G., Ng, B.W., Schoemann, A.M., et al. (2024). Power to Detect What? Considerations for Planning and Evaluating Sample Size. Personal. Soc. Psychol. Rev. *28*, 276–301. https://doi.org/10.1177/10888683241228328.
52. Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods *39*, 175–191.
53. Gong, X., Huang, Y.-X., Wang, Y., and Luo, Y.-J. (2011). Revision of the Chinese Facial Affective Picture System. Chin. Ment. Health J. *25*, 40–46.
54. Ekman, P., and Friesen, W.V. (1978). Facial Action Coding System: Investigator's Guide (Consulting Psychologists Press).
55. Ekman, R. (1997). What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS) (Oxford University Press).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
OSF: Raw and analyzed data	This paper	http://osf.io/xgfds
Dlib: Face landmark	King, 2009[24], http://dlib.net/files/shape_predictor_68_face_landmarks.dat.bz2	https://dlib.net/
Software and algorithms		
Python 3.9	https://www.python.org/	RRID: SCR_008394
NumPy 1.23	https://numpy.org/	RRID: SCR_008633
SciPy 1.9.1	https://scipy.org/	RRID: SCR_008058
Scikit-learn 1.2.2	https://scikit-learn.org/	RRID: SCR_002577
Dlib	King, 2009[24], https://dlib.net/	https://dlib.net/
Cili	Acland & Braver, 2014[50], https://github.com/beOn/cili	https://doi.org/10.5281/zenodo.48843
Original code	This paper	http://osf.io/xgfds

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact Lihui Wang (lihui.wang@sjtu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The raw data and analyzed data of this study are available at OSF, accession code: osf.io/xgfds. The datasets for detecting the face landmarks are available at: dlib.net/files/shape_predictor_68_face_landmarks.dat.bz2.
- All codes for reproducing the analyses reported in this study are available at OSF, accession code: osf.io/xgfds.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Participants

Twenty-nine university students (sex and gender: 18 females, 11 males, age: 23.2 ± 2.3 years old, ethnicity: Chinese) were recruited. One participant was excluded due to dropout, leaving 28 participants (17 females, 11 males, 23.1 ± 2.3 years old) included for data analysis. The sample was determined based on eye movement studies for face perceptions (e.g.,[6,8,14,17,18,29]), the sample size of which ranges from 20 to 30, and resource limitation. All participants reported normal or corrected-to-normal vision. Informed written consent was obtained from each of all participants prior to the experiment. The experiment was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the School of Psychological and Cognitive Sciences, Peking University (#2020-10-01). To evaluate the observed effect sizes, we performed the analysis of effect size sensitivity (suggested by the review[51]) using G-power 3.0.[52] Given the sample size (n = 28), the suggested power of 80%,[51] and the specified direction of our hypothesis concerning the function of Fix I and Fix II (e.g., one-tailed), the required Cohen's d is 0.48 for a meaningful effect. We thus identified the observed effects as achieving a large effect size given a $d > 0.48$ and a small effect size given $d < 0.48$.

METHOD DETAILS

Stimuli and apparatus

Stimuli were presented against a light gray background of a computer screen with a resolution of 1280 x 1024. Seventy grayscale images of faces were selected as the target faces from the Chinese Facial Affective Picture System.[53] The images of this dataset were standardized in the way that all faces were in frontal view, scaled to the same size, and hair and clothing were excluded. Seven emotion categories (angry, disgust,

fear, happy, neutral, sad, surprise) were included, with 10 images of different identities (5 females and 5 males) in each category. The ratings on the emotional category and intensity of the pictures can be seen in Table S1. Due to the limited number of identities available from the picture set, the identities between different emotional categories were not exclusively different.

Monocular eye movements were recorded using a video-based EyeLink 1000plus system (SR-Research, Canada) with a temporal resolution of 1000 Hz. A standard procedure of nine-point calibration and validation was performed at the beginning of each task, with a maximum error of 1.0° as the threshold.

Design and procedure

Participants were seated in a dimly-lit and sound-attenuated room at a viewing distance of 57 cm, with their heads positioned on a chin-rest. Each participant completed two tasks, an identity task and an emotion task, with the order of the two tasks counterbalanced across participants. In the identity task, participants were required to view a target face picture and then choose a face picture among 7 alternative face pictures that matched the identity of the target face. The 7 face pictures belonged to the same emotion category. In the emotion task, participants were required to view a target face picture and then choose a label that matched the emotion of the target face among 7 alternative labels.

Each trial began with a white cross in the center of the screen, followed by a green dot (radius = 5 pixels,0.15° of visual angle) randomly presented at one of the four corners of the screen (12.7° degrees of visual angle to the screen center). In 10% of all trials, a red dot (radius = 1 pixel, 0.03° of visual angle) was presented inside the green dot, and participants were instructed to press the space bar when they detected it. After the offset of the green dot, a target face picture (13.6 ° × 14 ° of visual angle, the eye-to-mouth distance around 3.5 ° × 6.5 ° degrees, i.e., 7.4° in total distance) was presented at the center of the screen for 1.5s (duration chosen according to 14), and participants were asked to view the face with free eye movements. Then, the 7 alternative options were presented, and participants were instructed to choose the correct answer by mouse click. In the identity task, the alternative options were 7 pictures of different identities (5.23 ° × 5.15 ° of visual angle), with only one of the same identities as the target picture. In the emotion task, the alternative options were 7 different labels of emotion words, with only one label matching the emotion of the target picture. The alternative options remained on the screen until a response was given. The inter-trial interval was 1 s.

Each task consisted of 4 blocks of equal length. In each block, each of the 70 pictures was presented as the target picture only once, rendering 280 trials in total (70 trials per block) for each task. There was a break between each two blocks. The order of task is randomized and counterbalanced across participants.

QUANTIFICATION AND STATISTICAL ANALYSIS

Extraction of fixations

Eye-movement data were extracted from the 1.5-s interval during which the target picture was presented. Data were firstly preprocessed using the *cili* module,[50] a python-based tool for detecting and correcting eye blinks. Fixations were then identified based on the velocity threshold of 30°/s and the acceleration threshold of 8000°/s^2. Fixations that occurred outside the target picture were excluded from analysis. Considering that the main purpose of the current study was to reveal the functions of the first two fixations, the analysis was focused on the first and the second fixations in each trial. The results of the fixations following are presented in Figure S2 and Figure S3.

K-means clustering of the fixations

The K-means clustering was performed to show the geographical distributions of the fixations. This method was to group the fixations into K clusters by minimizing the sum of Euclidean distances within the cluster, and the geometrical center of each cluster was used to represent the fixation tendency. The silhouette score[23] was introduced to determine the optimal K value, which measured how a fixation location was close to its own cluster than other clusters. For each of Fix I and Fix II and for each participant, we searched for an optimal K value ranging from 2 to 20 and identified the K value with the highest silhouette score, provided that the score exceeded the lowest score by more than 30%. If the highest score did not meet this threshold, the optimal K was set to 1. The cluster center with the optimal K value was mapped onto the face image. To visualize the cluster centers, a Gaussian kernel was applied to the two-dimensional distribution of centers using Kernel Density Estimation (KDE), with the bandwidth determined by Scott's Rule as a rule of thumb. A hundred contours were used to achieve a smooth demarcation between the areas hit by fixations and those missed by fixations.

Analysis of distances

For each face image, the facial landmarks were detected with Dlib[24] (https://github.com/davisking/dlib) and a trained dataset (http://dlib.net/files/shape_predictor_68_face_landmarks.dat.bz2). Dlib uses a combination of Histograms of Oriented Gradients (HOG) features and a supervised regression approach to predict the locations of these facial landmarks. Specifically, Dlib's facial landmark detector uses a cascaded ensemble of regressors called a shape predictor. This regressor ensemble is trained on a large dataset of annotated face images using a gradient descent optimization algorithm. Once trained, the shape predictor can efficiently and accurately detect facial landmarks in new images, even under various conditions (e.g., pose and illumination variations) by using the learned HOG features. The output of the Dlib facial landmark detection is a set of 68 2D facial landmarks. The 68 landmarks cover the contours of the whole face, the mouth (lips), the eyes and eyebrows, and the nose (see Figure 3A, left).

Four ROIs were defined: left eye, right eye, nose and mouth. The geometric center of all landmarks for a specific ROI was defined as the ROI center. For the two eyes, we used the 6 landmarks around the eyes. For the nose, we used the 4 landmarks along the nose bridge and the 5 landmarks along the nostril. For the mouth, we used the 14 landmarks around the contour of the lips, and the 4 landmarks inside the contour.

Two types of distance were calculated: the shortest distance (Shortest-Dis) and the sum of distances (Sum-Dis). The Shortest-Dis was defined as the Euclidean distance between a location on the face image and the closest ROI center, which quantified to which extent the location was closer to one ROI than the other ROIs (a lower distance indicates closer to the ROI). The Sum-Dis was defined as the sum of the Euclidean distances between a location on the face image and each ROI center, which quantified the cost of the location to get access to all four ROIs (a lower distance indicates a lower cost). To visualize each of the two types of distance on the face image, the distance was calculated for each pixel and averaged across all images. The heatmaps of the two types of distances are shown in Figure 3A (right). All calculations are based on the unit of visual angle.

For each participant and each of the two fixations, each type of distance was averaged across trials. For each type of distance, a paired-t test was performed to test the difference between Fix I and Fix II.

Feature-fixation correspondence

For each of the 70 target face images, the pixel-wise values were first normalized into Z scores. Then, the pairwise difference between a specific face image and each of the other face images was obtained by calculating the absolute value of the pixel-level difference between the Z scores. The pairwise absolute differences were averaged into a map which quantified to which extent a face image was different from the other images, and hence was identified as the map of Differentiating Features.

For Fix I and Fix II respectively, the fixation pattern was quantified by the pixel-level density map on the face image. For each target face in each task, the density map was obtained by the 2D histogram weighted by duration of the fixations across all trials of this face.

Cross-correlation was performed to investigate how the Differentiating Features correlate with the distributions of the first two fixations. Each pixel-wise feature map and each density map were flattened into a one-dimensional array respectively, and a cross-correlation was conducted between the vectorized feature values and the fixation density values. This correlation was performed separately between the Differentiating Feature and Fix I, and between the Differentiating Feature and Fix II. For each calculation, the fixation density vector was normalized by dividing the maximal density value, resulting in a normalized R value unbiased by the number of fixations located in the same pixel. The normalized r values were then transformed into *Fisher's Z* scores to normalize the underlying distribution, thereby enhancing the validity of subsequent t-tests. For each participant, the *Z*-scored r values were averaged across face images. A paired t-test was performed on the *Z*-scored r values to show if one of the first two fixations correlated more with the Differentiating Features than the other.

The cross-correlation was also performed for each emotion category. Considering the varied local features (e.g., 'action units' in facial expression studies[54]) in different face images may lead to misaligned features across stimuli within a specific emotion category, the face images were realigned prior to the calculation (see Figure S5). The face images within each category were realigned to achieve spatial consistency between the key features of individual faces within a category. For each face image, the 68 facial landmarks were first identified using Dlib. An affine transformation was then applied to align the landmarks using a set of reference landmarks derived from the average contour of all face images. A transformation matrix was calculated using the least-squares method to map a set of input landmarks to the reference landmarks. The transformation included scaling, rotation, translation, and shearing with the following steps: computing the geometric center of the input landmarks and the reference landmarks, scaling the input landmarks uniformly to the reference landmarks, applying translation to align them with the reference landmarks, and applying rotation to align the input landmarks with the reference landmarks. The resulting displacement vector was then added to the input landmarks to shift them to the corresponding positions in the reference face space. To ensure the correspondence between the fixation density map and the face image, the same procedure was also performed on the fixations. To keep the correspondence between the Differentiating Features and the fixation patterns, we applied the same transformation to the fixation locations. Note that here the stimuli alignment and fixation transformation were focused on emotional categories but not the image-based analysis above because the action units shared by the individual images within a certain emotion category may play a diagnostic role in recognizing the emotion category (e.g., Ekman[55]). Nevertheless, as a sanity check, we replicated the patterns of results when the transformation was taken into consideration for the image-based analysis (see Figure S6).

Then the feature-fixation correspondence of each fixation was compared between the correct and incorrect trials using paired *t* tests. To achieve equal statistical power for correct vs. incorrect trials (because there were more correct trials than incorrect trials), we applied a random sampling procedure to obtain 7 emotions x 28 participants x 2 trials (one correct and one incorrect) correlations and repeated this procedure 1000 times with replacement.

Prediction models of emotion categories

Three models were built to predict the emotion labels of the target face, with the fixation patterns included as the predictors. In model 1, the coordinates of Fix I (x_1, y_1) were included as predictors. In model 2, the coordinates of Fix II (x_2, y_2) were included as predictors. In model 3, the coordinates of both fixations (x_1, y_1, x_2, y_2) were included as predictors. For each prediction model, a random forest method with a 10-fold partitioning was used to train and cross-validate the model. For each participant, the procedure was repeated 100 times to calculate the mean prediction accuracy. For each model, the prediction accuracy was tested against chance-level (i.e., 1/7) using one-sample *t* test. The accuracies between model 2 and model 3 was compared using paired *t* test.