Humans Have Idiosyncratic and Task-specific Scanpaths for Judging Faces

Christopher Kanan

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Dina N.F. Bseiso

Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA

Nicholas A. Ray

Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA

Janet H. Hsiao

Department of Psychology, University of Hong Kong, Hong Kong

Garrison W. Cottrell

Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA

Abstract

Since Yarbus's seminal work, vision scientists have argued that our eye movement patterns differ depending upon our task. This has recently motivated the creation of multi-fixation pattern analysis algorithms that try to infer a person's task (or mental state) from their eye movements alone. Here, we introduce new algorithms for multi-fixation pattern analysis, and we use them to argue that people have scanpath routines for judging faces. We tested our methods on the eye movements of subjects as they made six distinct judgments about faces. We found that our algorithms could detect whether

Email addresses: ckanan@caltech.edu (Christopher Kanan), dbseiso@ucsd.edu (Dina N.F. Bseiso), niray@ucsd.edu (Nicholas A. Ray), jhsiao@hku.hk (Janet H. Hsiao), gary@ucsd.edu (Garrison W. Cottrell)

a participant is trying to distinguish angriness, happiness, trustworthiness, tiredness, attractiveness, or age. However, our algorithms were more accurate at inferring a subject's task when only trained on data from that subject than when trained on data gathered from other subjects, and we were able to infer the identity of our subjects using the same algorithms. These results suggest that 1) individuals have scanpath routines for judging faces, and that 2) these are diagnostic of that subject, but that 3) at least for the tasks we used, subjects do not converge on the same "ideal" scanpath pattern. Whether universal scanpath patterns exist for a task, we suggest, depends on the task's constraints and the level of expertise of the subject.

Keywords: eye movements, machine learning, scanpath routines, face perception

---

1. Introduction

Multi-Fixation Pattern Analysis (MFPA) is a new eye movement analysis technique that harnesses machine learning to make inferences about subjects from their eye movements (Benson et al., 2012; Greene et al., 2012; Tseng et al., 2013; Kanan et al., 2014; Borji and Itti, 2014). MFPA algorithms take a person's scanpath, a sequence of fixations, as their input and use the scanpath to infer traits such as the task the person was given. If an algorithm can make this inference above chance when trained on a person's scanpaths for specific tasks, then this suggests that the person might have scanpath routines for accomplishing one or more of the tasks. Prior work with MFPA has focused on validating the technique for inferring the task given to a subject when viewing scenes (Greene et al., 2012; Kanan et al., 2014; Borji and Itti, 2014) and for inferring whether the subject has a particular disease (Benson et al., 2012; Tseng et al., 2013). In this paper, we use MFPA to determine if people have scanpath routines for making different inferences about faces.

Humans make about three saccadic eye movements per second. Saccades are needed because the human retina has variable spatial resolution. It only acquires high resolution information in its central (foveal) region, with the resolution in the retinal periphery being far lower. The information in the periphery, along with information about the task being performed, can help direct saccades to diagnostic features for the task at hand. It makes sense, then, for humans to deploy scanpath routines for specific tasks so that diag-

nostic information can be acquired using as few fixations as possible. Formally, we define a scanpath routine as a task-specific sequence of fixations that exhibits a particular repeated spatial or spatio-temporal pattern. In order to rule out certain trivial cases, we also require that scanpath routines be acquired implicitly through learning, rather than elicited via direct instruction.

We hypothesized that scanpath routines for making common inferences about faces are likely to exist because some regions of the face are more diagnostic than others for some tasks. For example, the mouth is more diagnostic when judging whether a face is expressive or not and the eyes are crucial features for judging gender and identity (Gosselin and Schyns, 2001; Schyns et al., 2002). Similarly, in face recognition it has been shown that the left eye is the most diagnostic feature initially, followed by both eyes (Vinette et al., 2004). This finding is corroborated by results showing that people tend to fixate slightly to the left of the nose initially during face recognition (Hsiao and Cottrell, 2008). Taken together, these findings indicate that different face regions have varied diagnostic utility. However, the scanpaths people use to make inferences from faces may not be universal, because people have different experiences and slightly different visual systems. Peterson and Eckstein (2012) showed that the fixation points used by people to determine age, gender, and emotional state of a face differ across these three tasks. In subsequent work, they also showed that human eye movements during face identification were idiosyncratic (Peterson and Eckstein, 2013). Mehoudar et al. (2014) similarly found that scanpaths during face viewing are idiosyncratic, and that individuals continued to use the same idiosyncratic patterns when viewing faces 18 months later. Finally, several papers have shown that people's scanpaths have different properties when viewing novel faces compared to viewing familiar ones (Althoff and Cohen, 1999; Joyce, 2000).

The idea of scanpath routines is closely related to "scanpath theory" (Noton and Stark, 1971; Spitz et al., 1971). Scanpath theory argues that eye movements are generated in a top-down manner to facilitate correct recognition of an image by comparing it to previous experience. Learning a recognition task is taken to mean storing both the visual features and the motor sequence used to acquire the features. Recognition involves recapitulating the same scanpath when encountering the same stimulus. The strong form of scanpath theory predicts that individuals should deploy an identical pattern of eye movements during recognition, which is not consistent with

human behavior (Henderson, 2003). If humans did behave according to the strong theory, then it is likely that this behavior would have limited utility since humans rarely encounter exactly the same visual stimulus twice. A more general version of scanpath theory would predict that eye movements should be similar (statistically regular) between viewings of images from the same stimulus class, and this theory would allow scanpath routines to evolve over time to enable improved processing of the stimulus class, e.g., doing the task accurately using fewer fixations. This version of scanpath theory is consistent with our notion of scanpath routines.

To demonstrate the existence of scanpath routines for specific tasks, it is necessary to show that scanpaths are altered by the task, but this is not sufficient because it allows certain trivial cases. For example, Tatler et al. (2010) showed participants a photo of Alfred Yarbus wearing a coat and asked subjects various questions. When the subjects engaged in free viewing, the majority of their fixations were located on Yarbus's face, whereas when they were instructed to remember his clothing their fixations were more evenly distributed between his face and clothes. In their study, the instructions essentially required the subjects to view different parts of the image (the clothes). Obviously, one can easily create a situation where verbal instructions result in discriminable scanpaths. Trivially, one can ask the subject to look at the upper left hand corner of the image on one trial and the lower right on another. These examples are not scanpath routines because they are not acquired implicitly, i.e., the instruction tells the subject where to look.

For an observer to deploy a scanpath routine for a task, we believe two constraints must be met. First, the task needs to be one that an observer has experience with. Second, the task should be one in which the same task-diagnostic regions in each image will need to be fixated to perform the task accurately and using as few fixations as possible. In our study, these conditions are satisfied by using aligned facial images, such that the information is always in relatively the same locations in each image, and asking questions about them that subjects are likely to have experience with. From birth, humans acquire an enormous amount of experience in making judgments about others from their faces, suggesting that they will have established scanpath routines for efficiently answering these questions about faces. We ask our subjects to judge the age, fatigue, angriness, happiness, trustworthiness, and attractiveness of the people in the images. Because informative facial features are always located in the same relative position, it is possible to develop a scanpath routine so that task-relevant information can be obtained. To ad-

equately test for scanpath routines, we ask subjects the same six questions about every image, so that we are able to determine that the eye movements are not driven purely by the stimulus.

We attempt to find evidence for scanpath routines when making judgments about faces by using three different MFPA algorithms. The first method uses only summary statistics, i.e., the mean number of fixations in a trial and mean fixation duration, to make its inference. This approach ignores the spatio-temporal dynamics of the fixations, but it serves as a useful baseline. The second algorithm models the spatial distribution of fixations, including their duration at each location, but it ignores the temporal order information. The last algorithm is able to preserve information about the order of fixations as well as use spatial information. If any of the methods is above chance then we have strong evidence for scanpath routines for judging faces. By comparing the spatial and spatio-temporal methods, we can gain insight into the nature of these scanpath routines. For instance, if the spatio-temporal algorithm is significantly more accurate than the spatial algorithm then we can infer that there are diagnostic temporal regularities in the scanpaths.

We look for evidence of scanpath routines using a within-subjects and between-subjects analysis. If our algorithms are less accurate in a between-subjects analysis compared to within-subjects, then this suggests that people have idiosyncratic scanpath routines and that we can use our algorithms to infer the identity of the subjects.

2. Methods and Materials

2.1. Participants

The data used in our experiments is from 12 male and 12 female Caucasian UCSD students (mean age 19 years 8 months; age range 18-22), who received course credit for their participation. One additional female was recruited, but the data was excluded because she did not respond before the timeout in 98% of the trials. For the other participants, this occurred in 3% of trials on average (Min: 0%, Max: 11%). This data was not excluded in our analysis. All participants were right-handed based on the Edinburgh handedness inventory (Oldfield, 1971), and all had normal or corrected-to-normal vision. Participants gave informed consent after the study had been explained to them, the study was approved by UC San Diego's Institutional

Review Board, and the work was carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

2.2. Stimuli

The stimuli in our experiment consisted of 48, 561×701 pixel color face images, half female and half male. There were three images of each face model in this dataset, expressing either happy, angry, or neutral facial expressions. During a brief familiarization session to acquaint the participants with the experimental paradigm (described further in Section 2.4), five additional images of face models not used in the remainder of the study were employed. All face images were front-view, Caucasian, and none had facial hair or glasses. The images came from four age groups: child, young adult, adult, and elderly. Because no single dataset at the time of the study contained both images of children and elderly individuals, we combined images from two face datasets. Images of children came from the Radboud Faces Database (Langner et al., 2010). All of the other images came from the FACES dataset (Ebner et al., 2010), which contains images of young adults, adults, and elderly individuals.

Images were aligned without altering configural information by rotating, scaling, and translating them so that the triangle defined by the two eyes and the philtrum was as close as possible in euclidean distance to a reference triangle. This alignment was done using a nonreflective similarity transformation in MATLAB with the 'imtransform' function. To localize the face parts, we used a face part detection algorithm (Everingham et al., 2006). We then imposed a uniformly gray background. Participants viewed the stimuli on a 21 inch Sony CPD-G520 cathode ray tube monitor, with a refresh rate of 100 Hz, $\gamma = 2.2$, and a resolution of 1200x1024. The width of each face model's head, measured from the extreme outer edges of each pinna, on the screen was about 13 cm, and participant's viewing distance was 50 cm, so each head spanned about 15 degrees of visual angle. Example face model stimuli are shown in Figure 1.

2.3. Apparatus

An SR Research EyeLink II eye tracker was used to record participant eye movements. Binocular vision was used, but only the data from the eye with less calibration error was used in our analysis. The tracking mode was set to pupil only, with a sampling rate of 500 Hz. We used the EyeLink II's default algorithm to identify fixations.

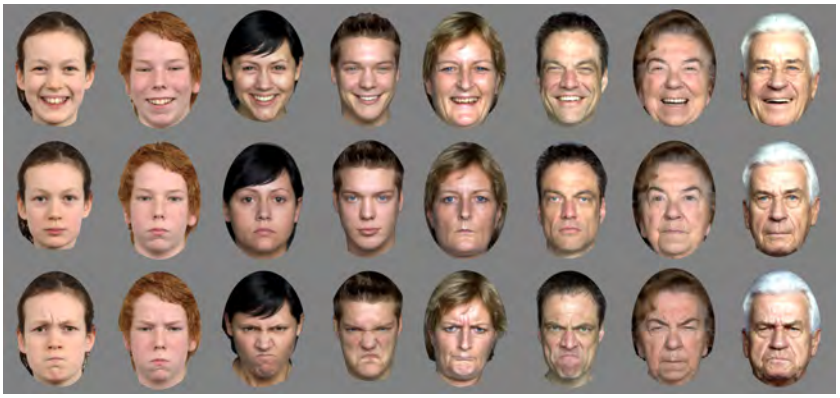

Figure 1: Example aligned face images used in our study. Each face model exhibits three emotional expressions: happy, neutral, and angry. There are four age groups: child, young adult, adult, and elderly.

An eight button Cedrus button box (four buttons per hand) was used to record participant responses. The right hand was used to record participant ratings of the stimuli, with each of the four buttons numbered sequentially. The left thumb was used to press the "GO" button, which was pressed by participants to proceed after viewing the instructions. This approach was used because the Cedrus button box has superior timing compared to the keyboard, it made it easier for participants to recall which key to press, and it reduced the likelihood that participants would look at their fingers during the experiment. All programming was done in MATLAB using Psychtoolbox-3 (Brainard, 1997) with the Eyelink toolbox extension (Cornelissen et al., 2002).

2.4. Design

A schematic of our experiment is given in Figure 2. The experimenter explained to the participant that they would see a face, and would need to rate it using a button box "as quickly and confidently as possible, because the image will timeout after a few seconds." The experimenter then familiarized the observer with the buttons of the input box and prompted the observer to begin the task by pressing a separate "GO" button. Participants were seated 50 cm away from the display monitor. After the initial eye tracker calibration, participants were familiarized with the experimental paradigm. Participants were asked to "Rank how clever this person is, on a scale from 1 (not clever) to 4 (very clever)." We restricted the ranking to only four choices because

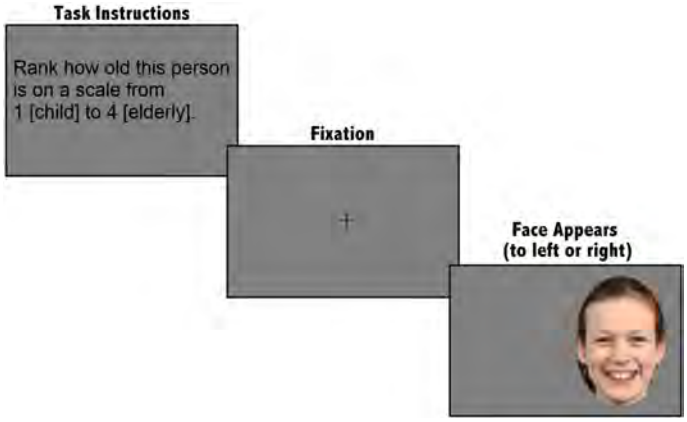

Figure 2: Our experiment was comprised of six blocks. In each trial, participants were instructed with the block's task and then one of the faces appeared to the right or left of the fixation cross. Each face was observed only once in each block.

the button box only had four buttons for the right hand. Subsequently, a fixation cross appeared in the center of the screen. The experimenter then initiated the trial when the participant was gazing at the cross, causing the fixation cross to disappear and the image to appear randomly either on the right or left side of the screen. The distance on the screen from the initial fixation point to the nasion was 11.35 cm; hence, a 13 degree saccade was needed to fixate the center of the face. Once the participant had ranked the image according to the task, or 3 seconds had elapsed, the on-screen prompt returned, followed by gaze correction, and then the next image would appear. The block continued in this manner until all five images had been viewed by the observer and the block was completed. This familiarization block used five images that were distinct from the 48 images used in each block of the main experiment. The data collected from this familiarization block was excluded from analysis.

After familiarization, participants viewed the 48 images in 6 blocks, so each image was seen 6 times. Each block had a different prompt, and a Latin square design was used to determine the order of the blocks for each participant. The Latin square design was used to compensate for any potential memory confound due to participants seeing the faces multiple times in our between-subjects analysis. The order of the images was randomized within each block. The six block prompts were (1) Rank how old this person is on a scale from 1 [child] to 4 [elderly]; (2) Rank how fatigued this person is on

a scale from 1 [alert] to 4 [tired]; (3) Rank how happy this person is, on a scale from 1 [not happy] to 4 [very happy]; (4) Rank how angry this person is, on a scale from 1 [not angry] to 4 [very angry]; (5) Rank how trustworthy this person is, on a scale from 1 [not trustworthy] to 4 [very trustworthy]; and (6) Rank how attractive this person is, on a scale from 1 [not attractive] to 4 [very attractive]. The rest of the experiment was identical to the earlier described familiarization block.

The "age" task is the only one with a "ground truth" answer, based on the four categories of facial images that were used (child, young adult, adult, elderly). This task allowed us to check that the subjects were engaged in their tasks since we know the correct response. The other tasks are ones that humans often engage in as social animals. Previous work has shown that judgments of emotion (Ekman, 1973; Izard, 1971), fatigue (Sundelin et al., 2013), trustworthiness and attractiveness (Willis and Todorov, 2006) are reliably assessed by observers. There are also subtle differences in observer's eye movements during face perception when judging age and fatigue (Nguyen et al., 2009), suggesting that scanpath routines could be used in these tasks.

2.5. Algorithms.

Most classifiers in machine learning require all input feature vectors to have a fixed-dimensionality. However, the number of fixations acquired in each trial is variable-length. The main challenge in MFPA is turning a trial's fixation features into a single fixed-dimensionality vector that captures diagnostic information. We implemented three primary algorithms for turning a trial's scanpath features into a fixed-dimensionality feature vector. An overview of them is given in Figure 3. We also used four additional combinations formed from those algorithms. The first method is based on using summary statistics from each trial, and it serves as a baseline. The second uses the spatial characteristics of the fixations, and the third uses the spatio-temporal characteristics of the scanpath. The latter two methods are both based on Fisher vectors. The same classification algorithm is then applied to each of these feature representations.

During each trial we acquire a sequence of fixation features. For a fixation t these consist of the (x_t, y_t) screen coordinates of the location fixated and the duration d_t of the fixation. The data is preprocessed by removing the first fixation, which is generally at the location of the fixation cross, followed by centering the fixations onto the stimulus' screen location. Trials in which only a single fixation was recorded were discarded, which occurred in less

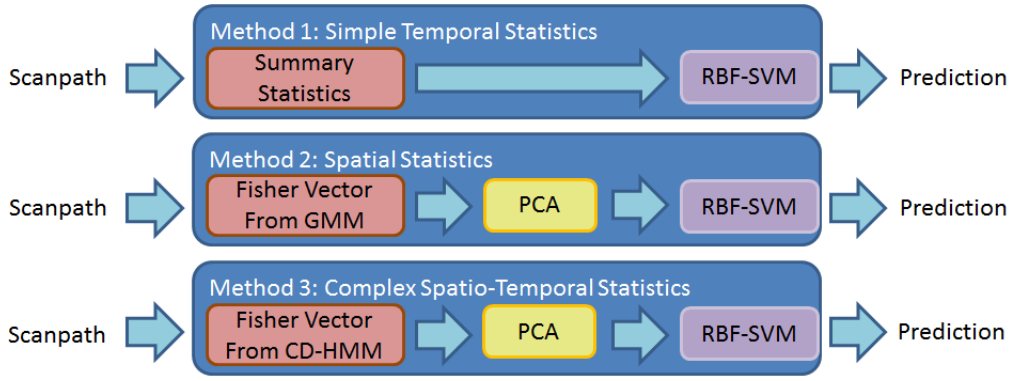

Figure 3: An overview of the three three algorithms that we use to turn variable-length scanpaths into fixed-length vectors that can be used for classification. The summary statistics method only captures simple temporal statistics about the scanpath. The Fisher vector method using a GMM captures spatial statistics about where the participant is looking and for how long. The Fisher vector method using a CD-HMM goes further by also capturing more complex spatio-temporal information, such as the transitions from one fixation region to the next. For the Fisher vector methods, PCA is used for dimensionality reduction. The same classification algorithm, a support vector machine using an RBF kernel, is used for all three feature representations.

than 0.5% of all trials. This data served as input to each of the MFPA methods. In a trial with T fixations, we represent its information using the matrix

$$\mathbf{X}_{trial} = \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_T \end{array} \right] = \left[\begin{array}{cccc} x_1 & x_2 & & x_T \\ y_1 & y_2 & \cdots & y_T \\ d_1 & d_2 & & d_T \end{array} \right],$$

where each column represents the screen gaze coordinates and fixation duration. We will refer to this representation in our description of the three primary methods we use to turn $\mathbf{X}_{trial}$ into a fixed-dimensional representation that can be used with an off-the-shelf classifier.

The summary statistics method turns a trial's fixations into a 2-dimensional feature vector containing the number of fixations in the trial and the mean fixation duration:

$$\Phi_{SS}\left(\mathbf{X}_{trial}\right) = \left[\begin{array}{cc} T & \frac{1}{T}\sum_{t=1}^{T} d_t \end{array} \right]^T.$$

Each dimension of this representation is then normalized by dividing by the standard deviation of the training data. This approach is similar to the method used by Greene et al. (2012) and later by Kanan et al. (2014).

The summary statistics algorithm only captures the simplest temporal statistics, and it ignores all spatial properties and more complex temporal characteristics. The next two methods for turning a trial's time-series features into a fixed-dimensionality vector can retain the spatial or spatio-temporal aspects, and they are both based on the idea of Fisher vectors (Jaakkola and Haussler, 1998; Perronnin et al., 2010). To use Fisher vectors, a parametric generative model $p(\mathbf{X}|\mathbf{\Theta})$ is trained, where $\mathbf{X}$ is the training data and Θ are the parameters of the model that have been estimated using maximum likelihood estimation. A trial's time-series features $\mathbf{X}_{trial}$ are turned into a Fisher vector Φ_{FV} by examining how they would alter the maximum likelihood parameter estimate:

$$\Phi_{FV}\left(\mathbf{X}_{trial}\right) = \nabla_{\Theta} \log p\left(\mathbf{X}_{trial}|\Theta\right).$$

The dimensionality of this representation depends only on the number of parameters in Θ, and it is invariant with respect to the length of the time-series. The idea behind Fisher vectors is that the gradients for two trials from the same category will be similar.

Before using them as input to a classifier, the Fisher vector features are normalized in a two-step process that has been shown to be crucial for achieving state-of-the-art performance with them (Perronnin et al., 2010). The first step is to take a sign-preserving square root of the features, i.e., $f\left(z\right) = \text{sign}\left(z\right)\sqrt{|z|}$ is applied element-wise to the features. The second step is to make the features unit length by dividing by their Euclidean norm.

We use two different generative models with Fisher vectors. We briefly summarize how to compute them here, but see Perronnin et al. (2010) and van der Maaten (2011) for further details. The first method is a Gaussian mixture model (GMM), which will represent the spatial characteristics of a scanpath without preserving any of its temporal properties. This Fisher vector representation has been very successful at object (Perronnin et al., 2010) and face (Simoyan et al., 2013) recognition problems in computer vision; we are the first to apply it to eye movement data. To compute these Fisher vectors, we used the implementation in the MATLAB VLFeat toolbox (Vedaldi and Fulkerson, 2008), which uses Gaussians with diagonal covariance. The parameters used to construct Fisher vectors were the means and covariances of the Gaussians. The mixture weight Fisher vectors are typically not discriminative and were not supported by the toolbox. Formally, a GMM composed

of K Gaussians is given by the equation,

$$p\left(\mathbf{x}\right) = \sum_{k=1}^{K} w_k \mathcal{N}\left(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k\right)$$

where w_k are the mixture weights and $\mathcal{N}\left(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k\right)$ are the Gaussian densities in the mixture, which each have their own mean μ_k and diagonal covariance $\mathbf{\Sigma}_k$. The parameters of the GMM are fit to the training data. Example GMMs are shown in Figure 4. As is standard with Fisher vectors, only one GMM is learned, regardless of the number of categories that need to be discriminated. The gradients of the means and covariances are used to generate the GMM Fisher vectors that summarize the information in a trial. For simplicity and to be consistent with other recent papers, e.g., Simoyan et al. (2013), we drop the vector differentials and just give the equations for computing the Fisher vector features. For each mixture component k, these are given for the means by

$$\mathbf{u}_k = \frac{1}{T\sqrt{w_k}} \sum_{t=1}^{T} q_{tk} \mathbf{\Sigma}_k^{-1} \left(\mathbf{x}_t - \mu_k\right)$$

and for the covariances by

$$\mathbf{v}_k = \frac{1}{T\sqrt{2w_k}} \sum_{t=1}^{T} q_{tk} \left[\left(\mathbf{\Sigma}_k^{-1} \left(\mathbf{x}_t - \mu_k\right)\right) \circ \left(\mathbf{\Sigma}_k^{-1} \left(\mathbf{x}_t - \mu_k\right)\right) - \mathbf{1}\right],$$

where T is the number of fixation features in the trial, $\circ$ represents the element-wise matrix product (i.e., Hadamard product), $\mathbf{1}$ is a vector of ones, and

$$q_{tk} = \frac{\exp\left[-\frac{1}{2}\left(\mathbf{x}_t - \mu_k\right)^T \mathbf{\Sigma}_k^{-1} \left(\mathbf{x}_t - \mu_k\right)\right]}{\sum_{j=1}^{K} \exp\left[-\frac{1}{2}\left(\mathbf{x}_t - \mu_j\right)^T \mathbf{\Sigma}_j^{-1} \left(\mathbf{x}_t - \mu_j\right)\right]},$$

is the assignment strength the GMM gives the fixation features $\mathbf{x}_t$ to each mixture component. The unnormalized GMM Fisher vector for a trial is then created by concatenating the $\mathbf{u}_k$ and $\mathbf{v}_k$ vectors:

$$\Phi_{GMM}\left(\mathbf{X}_{trial}\right) = \left[\mathbf{u}_1, \mathbf{v}_1, \ldots, \mathbf{u}_K, \mathbf{v}_K\right]^T.$$

This representation is then normalized using the two-step procedure described earlier.

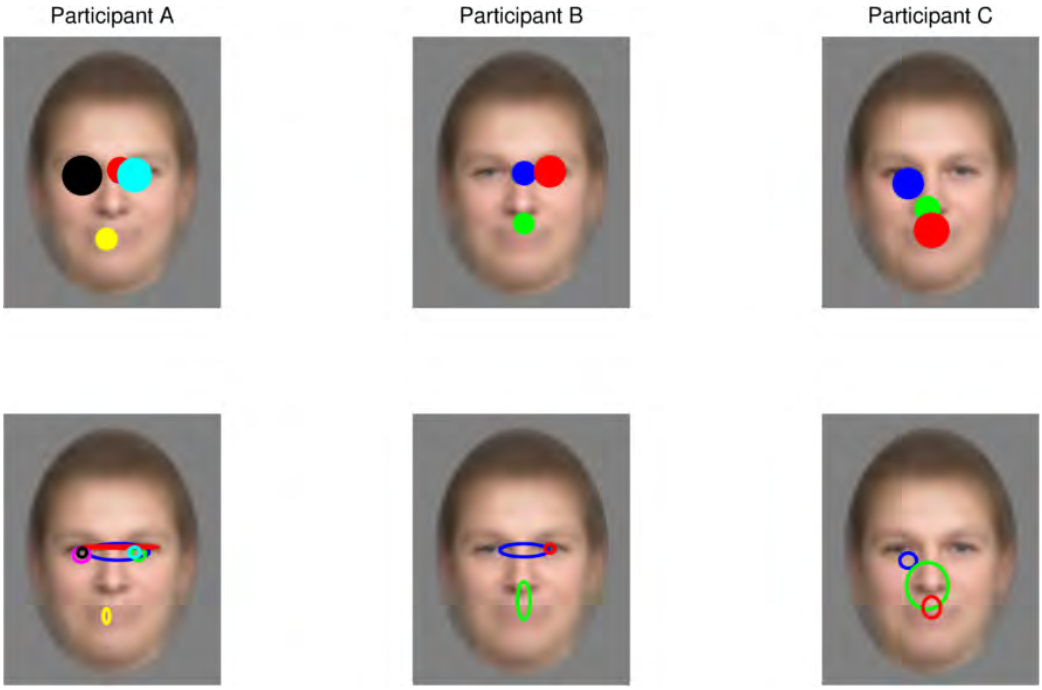

Figure 4: Example GMMs trained on the fixations of three participants shown on an average face. The top row shows the location of each model's Gaussians, with the Gaussian's duration parameter represented by the size of the circle. The bottom row shows the covariances of the Gaussians' spatial parameters.

The second generative model we use to compute Fisher vectors is a continuous density hidden Markov model (CD-HMM), i.e., an HMM with Gaussian emissions. This method will model both the spatial and temporal properties of a trial's scanpath features. A CD-HMM has one Gaussian per hidden state. Fisher vectors with HMMs and CD-HMMs have been used in bioinformatics for protein classification (Jaakkola et al., 2000), in computer vision for activity recognition (Sun and Nevatia, 2013), and in speaker identification (Wan and Renals, 2002). For the CD-HMM Fisher vector model, we used the MATLAB CD-HMM implementation provided by van der Maaten (2011), and we set it to use diagonal covariance matrices. For the CD-HMM model, the parameters used to construct Fisher vectors were the means and covariances of the Gaussians and the state transition matrix. Formally, a CD-HMM models the joint distribution of a trial's fixation features $\mathbf{X}$ over

sequences of hidden states $\mathbf{s} = \{s_1, s_2, \ldots, s_T\}$ and is given by

$$p(\mathbf{X}, \mathbf{s}) = p(s_1) \prod_{t=1}^{T-1} p(s_{t+1}|s_t) \prod_{t=1}^{T} \mathcal{N}\left(\mathbf{x}_t|\mu_{s_t}, \mathbf{\Sigma}_{s_t}\right),$$

where $p(s_1)$ is the initial hidden state distribution, $p(s_{t+1}|s_t)$ are the state transition probabilities, and $\mathcal{N}\left(\mathbf{x}_t|\mu_{s_t}, \mathbf{\Sigma}_{s_t}\right)$ are the Gaussian emission densities associated with each hidden state. In CD-HMM Fisher vectors, only one CD-HMM is learned, regardless of the number of categories that need to be discriminated. This is because trials belonging to the same category will presumably change the model in similar ways. CD-HMMs trained on the fixations of three participants are shown in Figure 5. For a CD-HMM with K states, a trial's Fisher vectors for the means of the Gaussians are given by

$$\mathbf{u}_k = \frac{1}{T} \sum_{t=1}^{T} \gamma_{tk} \Sigma_k^{-1} (\mathbf{x}_t - \mu_k),$$

the Fisher vectors for the covariances by

$$\mathbf{v}_k = \frac{1}{2T} \sum_{t=1}^{T} \gamma_{tk} \left[\left(\mathbf{\Sigma}_k^{-1} (\mathbf{x}_t - \mu_k)\right) \circ \left(\mathbf{\Sigma}_k^{-1} (\mathbf{x}_t - \mu_k)\right) - \mathbf{\Sigma}_k^{-1}\mathbf{1}\right],$$

and the elements of the Fisher vector for the state transition matrix by

$$h_{kj} = \frac{m_{kj}}{a_{kj}},$$

where γ_{tk} is the CD-HMM posterior probability over the states given the observations, m_{kj} is the CD-HMM posterior probability over the transition edges, and a_{kj} is the element of the state transition matrix representing the probability of transitioning from state k to state j. All of these are then concatenated to form the CD-HMM Fisher vector representation:

$$\Phi_{HMM}(\mathbf{X}_{trial}) = [\mathbf{u}_1, \mathbf{v}_1, \ldots \mathbf{u}_K, \mathbf{v}_K, \mathbf{h}]^T,$$

where $\mathbf{h}$ is all of the h_{kj} elements concatenated into a vector. Subsequently, we apply the two-step normalization process described earlier to the CD-HMM Fisher vector.

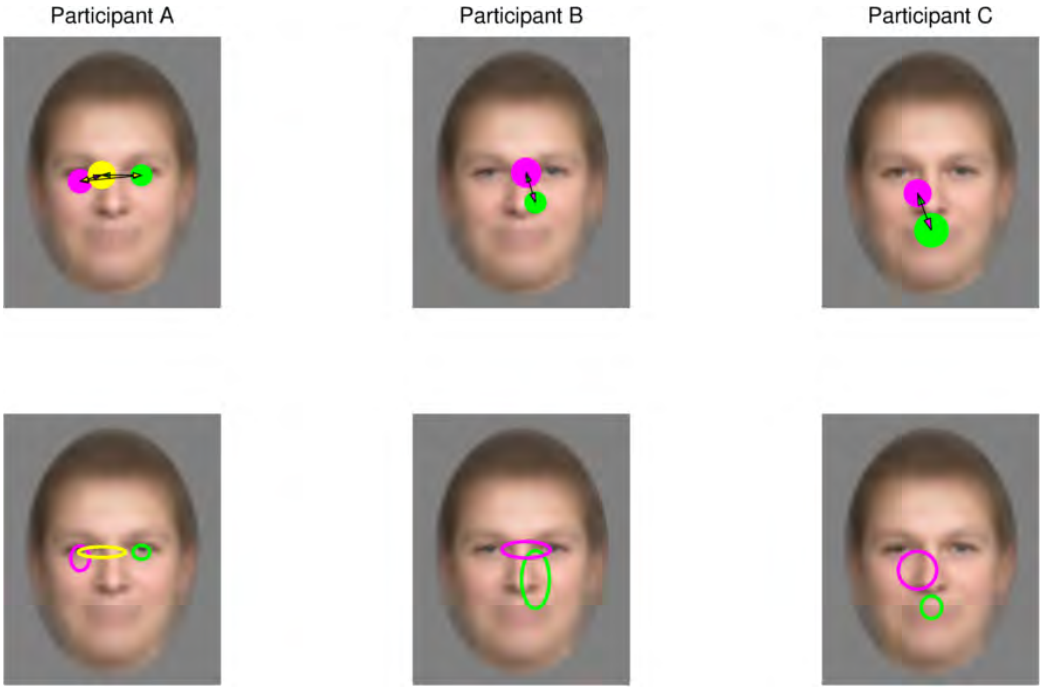

Figure 5: Example CD-HMMs trained on the fixations of three participants shown on an average face. For visualization purposes, we limited the CD-HMMs to at most three states. The top row shows the location of each model's Gaussians, with the Gaussian's duration parameter represented by the size of the circle and the transition matrix is shown by arrows (ignoring self-transitions and the strength of the transitions). The bottom row shows the covariances of the Gaussians' spatial parameters.

In addition to these three primary feature types (summary statistics, CD-HMM Fisher vectors, and GMM Fisher vectors), we used all four combinations of them by concatenating them together: summary statistics with GMM Fisher vectors, summary statistics with CD-HMM Fisher vectors, GMM Fisher vectors with CD-HMM feature vectors, and summary statistics with GMM Fisher vectors and CD-HMM feature vectors. By comparing these models we can measure the amount of improvement gained by incorporating spatial or temporal properties. If the CD-HMM is superior to the other models or increases their accuracy when combined with them, then this suggests the temporal order of the fixations provides some diagnostic information.

Across all of these representations of a trial's features, the same radial-basis function support vector machine (SVM) classification algorithm was

used. We used the LIBSVM toolbox (Chang and Lin, 2011). For all methods except the summary statistics algorithm, we first reduced the dimensionality of the feature vectors using whitened principal component analysis (PCA). While the number of the Gaussians and states influences the dimensionality of the Fisher vector representations, PCA plays a complimentary role that has an independent influence on classification accuracy. The width of the radial-basis functions, SVM cost parameter, number of principal components, number of clusters (for the GMM model), and number of hidden states (for the CD-HMM model) were tuned with 4-fold cross validation using the training data, with the width and cost parameters chosen from $2^{-8}, 2^{-7}, \ldots, 2^{8}$, number of principal components chosen from $2^{0}, 2^{1}, \ldots, 2^{8}$ (or fewer, depending on the dimensionality of the Fisher vector representation), and number of clusters or hidden states chosen from $1, 2, \ldots, 10$.

Our data is slightly unbalanced due to trials with only one fixation being excluded, as described earlier. Because of this, we use a random classifier that makes random predictions from a uniform distribution to calculate chance performance in each experiment. If no trials were dropped, then in our task prediction experiments chance would be 16.67% and in our participant identity prediction experiment chance would be 4.17%. In all experiments, the random classifier is very close to these levels.

3. Results

Because we know the age category for all of the stimuli, we can use the age ranking task to assess how engaged our subjects were during the experiment. We found that participants accurately predicted the age category, with a mean absolute error of 0.19 (95% CI = 0.14–0.24) in their prediction of the age group. Since age was ranked from 1 to 4, the maximum possible mean absolute error is 3, and if participants were randomly guessing it would be 1.25 on average. In trials where subjects incorrectly predicted the age, they were off by no more than one age category in 99.1% of trials (e.g., misclassifying a young adult as an adult). These statistics suggest our participants were genuinely trying to do the task. The relative frequency of the ratings for each task is shown in Figure 6. Participant agreement on the task ratings, mean number of fixations per trial, and mean reaction time per trial are shown in Table 1. Since the number of fixations varies among the tasks (e.g., age vs. trustworthy), this suggests that the summary statistics algorithm will be sufficient to perform above chance in a between-subjects analysis. Table

Table 1: Participant agreement computing using the intraclass correlation coefficient, the mean number of fixations, and the mean reaction times for each task across all participants. A 95% confidence interval is given for the number of fixations and reaction times. The tasks in which participant agreement was higher generally required fewer fixations than for those tasks where it was lower.

	Agreement	Number of Fixations	Reaction Times (s)
Age	0.85	3.05 ± 0.10	1.20 ± 0.03
Happy	0.75	3.59 ± 0.11	1.44 ± 0.04
Angry	0.65	3.67 ± 0.11	1.46 ± 0.03
Tired	0.38	4.29 ± 0.11	1.71 ± 0.03
Trustworthy	0.31	4.23 ± 0.12	1.76 ± 0.04
Attractive	0.30	4.03 ± 0.12	1.56 ± 0.03

Table 2: The frequency of fixations to each face region across participants for each of the six tasks.

	Left Eye	Right Eye	Nose	Mouth
Overall	27.3%	26.1%	30.4%	16.2%
Happy	26.4%	24.6%	29.9%	19.1%
Angry	24.2%	25.9%	30.9%	19.0%
Tired	28.3%	27.0%	29.0%	15.7%
Attractive	28.2%	25.4%	31.6%	14.8%
Trustworthy	29.2%	26.4%	29.7%	14.7%
Age	26.4%	27.3%	32.0%	14.3%

2 shows the frequency of fixations to each face region across participants. The statistics across tasks differ little, suggesting that a between-subjects approach using spatial statistics will perform poorly.

Figure 7 shows the density of participant fixation locations across the six tasks. To generate the density plots, we used a Gaussian kernel density estimation method in which the bandwidth is automatically estimated (Botev, 2006). When the fixations from all participants are combined, we get the usual "T" shaped pattern of fixations around the eyes and mouth for all of the tasks. However, there are strong qualitative individual differences among participants, consistent with the findings of others (Peterson and Eckstein, 2013; Mehoudar et al., 2014), suggesting that we will be able to infer participant identity. For each participant, the differences in fixation density across tasks are subtle. This suggests that inferring the task will be difficult for the algorithms if restricted to only the gaze coordinates.

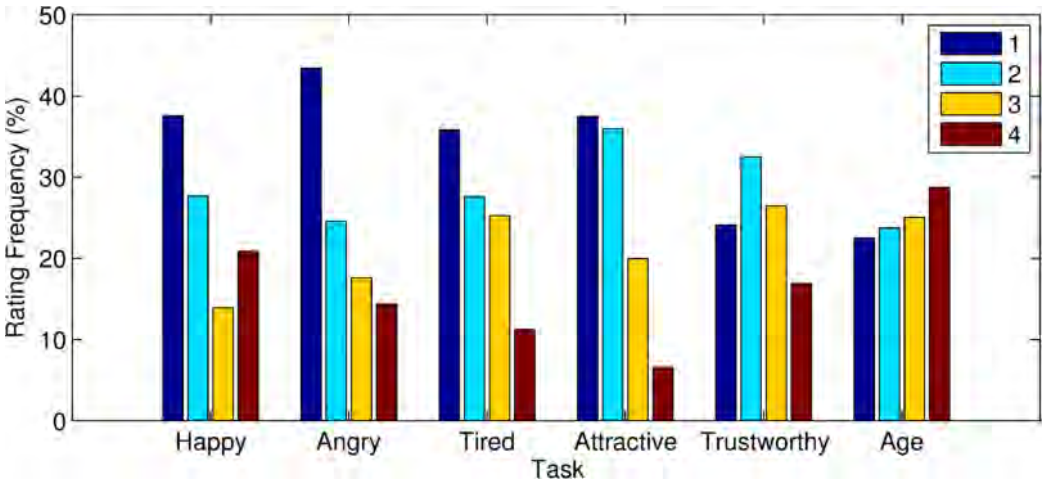

Figure 6: The relative frequency of the four ratings picked by all participants across the six tasks. Age was closest to a uniform distribution, which is expected since participants classified age with high accuracy and each age category contained the same number of face models. Attractive is the farthest from uniform, with participants generally rating our face models as unattractive.

3.1. Within-Subject Task Prediction

Our first task prediction experiments are within-subject. If we can judge the task a subject is performing based on their scanpaths, this would suggest that they have a scanpath routine for the task. For each subject we use leave-one-out cross validation, i.e., we train all of the methods on 287 trials, test on the remaining hold-out trial, and repeat this process 288 times. We also generated labels at random (denoted Random Classifier) to calculate chance, since the number of trials are slightly unbalanced due to dropped trials. As shown in Table 3, all of the algorithms performed above chance, with the summary statistics algorithm performing worst and the combination of all three methods performing best. This result indicates that it is possible to discern a participant's face judgment task at above chance levels solely from their scanpath.

What role does temporal information play in scanpath routines? Since the summary statistics model is above chance, this means that simple temporal regularities exist and are diagnostic. Does incorporating the order of fixations provide any benefit over a spatial model combined with simple temporal statistics? While there was no significant difference between the CD-HMM method compared to the GMM method with summary statistics, $t(6877) =$

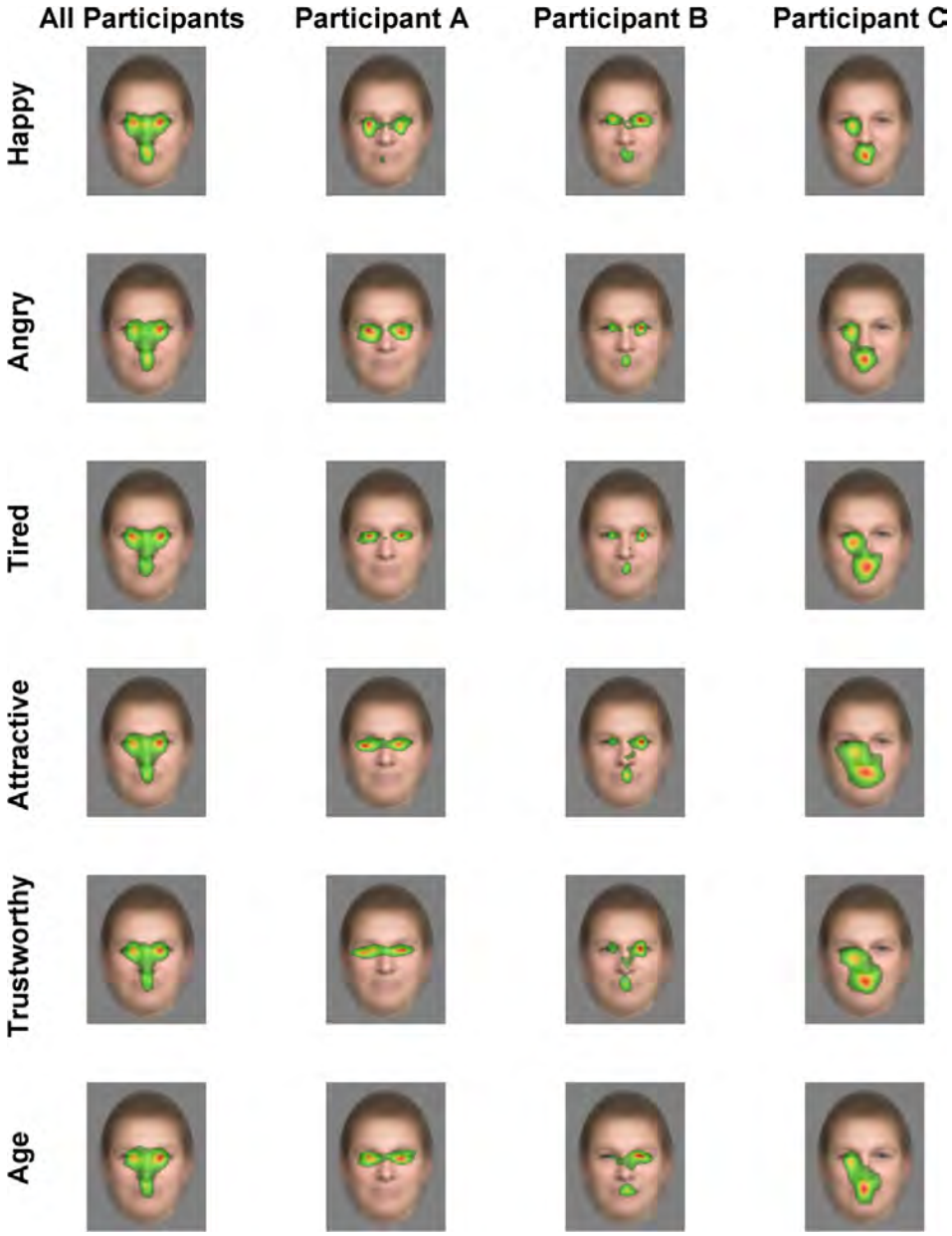

Figure 7: The distribution of fixations across the six tasks for all participants (first column) and for three particular participants (remaining columns). Qualitatively, there are strong individual differences among participants, e.g., participant C primarily looks at the left eye and mouth whereas participant A primarily looks at both eyes and rarely the mouth.

Table 3: Mean predictive accuracy and 95% confidence intervals from within-subject experiments for all methods. SS denotes summary statistics method, GMM denotes the GMM Fisher vector model, and CD-HMM denotes the CD-HMM Fisher vector method.

Method	Mean (%)	95% CI
Random Classifier	16.67	15.79–17.56
SS	26.00	24.96–27.05
GMM	31.23	30.14–32.34
CD-HMM	32.41	31.30–33.53
SS + GMM	33.48	32.37–34.61
SS + CD-HMM	33.91	32.79–35.04
GMM + CD-HMM	33.99	32.87–35.13
SS + GMM + CD-HMM	36.30	35.17–37.45

$1.62, p = 0.10$, we did find that the combination of all three methods was significantly more accurate than the other models (see Table 3). Moreover, adding the CD-HMM method to any other method improved accuracy. Taken together, these results indicate that complex temporal statistics, including the order of fixations, are incorporated into scanpath routines to a limited extent since they are of diagnostic value.

We have argued that, because accuracy is above chance, scanpath routines for judging faces exist. Do we have scanpath routines for all of the tasks in our study or just a subset of them? We can gain some insight by analyzing the confusion matrices for the MFPA methods, which are shown in Figure 8. All of the methods were above chance for all of the tasks, except for angry with the summary statistics method (16% correct). For the combination of all three MFPA methods, the age task had the highest accuracy (41% correct) and angry had the lowest (31% correct). Tasks that are confused with greater frequency by the algorithms indicate that the scanpath statistics for the these tasks are more similar than other tasks. For the combination of all three MFPA methods, this occurred most between happy and angry as well as attractive and trustworthy. The happy/angry confusion could indicate that people have a general scanpath routine for classifying emotional facial expressions, potentially involving more frequently looking at the mouth than for other tasks (see Table 2). Several studies have shown that human judgments of appearance and trustworthiness are correlated (Budesheim and DePaola, 1994; Zaidel et al., 2003; Kleisner et al., 2013), and this may be the reason why the scanpaths for attractiveness and trustworthiness are more

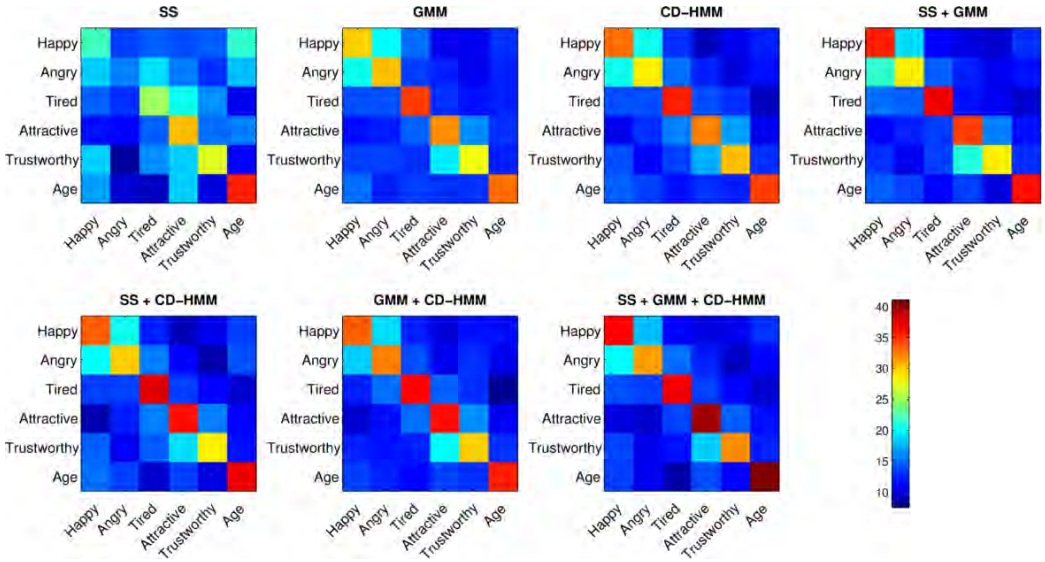

Figure 8: Task inference confusion matrices from the within-subjects analysis for all methods. Each column indicates the predicted category and each row represents the actual class. SS denotes summary statistics method, GMM denotes the GMM Fisher vector model, and CD-HMM denotes the CD-HMM Fisher vector method.

frequently confused by the algorithms.

The confusion matrices shown in Figure 8 are averaged across all participants, but how similar are the confusion matrices computed for each of the 24 participants? To measure this per algorithm, we treated the task confusion matrix computed for each participant as a vector and then calculated the intraclass correlation coefficient. The intraclass correlation coefficient was moderate (0.76) for the summary statistics method, but was high (0.94–0.96) for all of the other algorithms. This indicates that when the algorithms are trained on different subjects, they still make the same kinds of confusions.

3.2. Between-Subject Task Prediction

If subjects have scanpath routines for judging faces, do they have the same scanpath routines across subjects? To the extent that this is the case, we should be able to tell what a subject is doing from another subject's scanpath data. To test this hypothesis, we trained the algorithms on all of the data from 23 of the 24 participants, and tested on the remaining hold-out participant. This was repeated 24 times, with each participant serving as a hold-out. This approach uses 23 times more training data than our

Table 4: Mean predictive accuracy and 95% confidence intervals from the between-subject experiments for all methods. SS denotes summary statistics method, GMM denotes the GMM Fisher vector model, and CD-HMM denotes the CD-HMM Fisher vector method.

Method	Mean (%)	95% CI
Random Classifier	16.68	15.80–17.58
SS	22.16	21.18–23.16
GMM	18.19	17.28–19.12
CD-HMM	20.38	19.44–21.36
SS + GMM	19.85	18.91–20.81
SS + CD-HMM	20.94	19.98–21.92
GMM + CD-HMM	20.66	19.71–21.64
SS + GMM + CD-HMM	21.39	20.42–22.38

within-subjects experiments, so if people have universal scanpath routines for these tasks we would expect performance to be at least as good as our within-subjects results. As shown in Table 4, all of the algorithms performed above chance; however, the results are worse than the within-subjects results. The best method in the between-subjects analysis is 39% worse, in relative terms, than the best method in the within-subjects analysis. Also, unlike the within-subjects analysis, no method is better than using the summary statistics method. Moreover, when the summary statistics method is combined with either the GMM or CD-HMM methods, accuracy is reduced, suggesting they are adding noise to the classifier's input. These results indicate that scanpath routines for judging faces are idiosyncratic because spatial and complex spatio-temporal information provides no benefit over simple summary statistics. The GMM Fisher vector model performs worst, suggesting that the spatial distribution of fixations differs among participants. The CD-HMM model performs significantly better than the GMM method, and this is likely because the CD-HMM method is implicitly capturing some of the low-level summary statistics.

3.3. Participant Identity Prediction

Since our within-subjects results are significantly higher than our between-subjects results, this indicates that people have idiosyncratic scanpaths. To verify this, we trained classifiers to infer participant identity, i.e., the algorithms were trained on labeled scanpaths from each of the 24 subjects and given an unlabeled scanpath they would predict which subject generated it.

Table 5: Mean predictive accuracy and 95% confidence intervals from participant prediction experiments for all methods. SS denotes summary statistics method, GMM denotes the GMM Fisher vector model, and CD-HMM denotes the CD-HMM Fisher vector method.

Method	Mean (%)	95% CI
Random Classifier	4.18	4.02–4.35
SS	14.40	14.11–14.69
GMM	52.78	52.37–53.20
CD-HMM	58.64	58.23–59.05
SS + GMM	53.76	53.35–54.18
SS + CD-HMM	58.16	57.75–58.57
GMM + CD-HMM	60.96	60.55–61.36
SS + GMM + CD-HMM	61.79	61.39–62.19

For this experiment, we trained the classifiers using 240 randomly selected trials per participant (40 per task), and we tested on the remaining trials. Our results are calculated over 50 random cross-validation runs. As shown in Table 5, all algorithms performed above chance, with the summary statistics algorithm performing worst and the combination of all three methods performing best. While the CD-HMM and GMM Fisher vector methods were comparable in our within-subjects analysis, the CD-HMM method achieved significantly higher accuracy at participant identification. This indicates that the temporal regularities in a subject's scanpaths are indicative of their identity.

4. Discussion

The experiments we conducted were designed to determine if it was possible to infer which face inference task a subject was trying to accomplish, solely from their eye movements. All of the tasks we gave the subjects were detectable from their scanpaths at a level well above chance. We conclude from these results that individual humans have scanpath routines for faces to answer particular questions. However, we cannot conclude that subjects were using the *same* scanpath routines for three reasons: (1) performance was much worse in trying to detect a subject's task based upon all of the other subjects' data, despite using a larger amount of training data, (2) unlike our within-subject results, no algorithm outperformed the summary statistics method in our between-subject analysis, which means that capturing

more complex spatio-temporal properties was not helpful, and (3) we were able to accurately predict participant identity, which should work poorly if participants are using the same scanpath routines. This finding is consistent with other reports that human scanpaths are idiosyncratic (Noton and Stark, 1971; Foulsham et al., 2012). While we believe it is possible to develop superior algorithms for MFPA, we do not think that doing so will change our conclusion that people have scanpath routines for judging faces since this conclusion is based on the algorithms performing above chance.

We used the Cartesian coordinates of our subject's fixations as one of the features given to the classifiers. One of the reasons why this was effective is that we implicitly used an object-centered coordinate system because we aligned the face stimuli, so the gaze coordinates contain information about what facial parts are being fixated. We suspect that performance would be much poorer if the images were unconstrained. One way to get around this is to use an explicitly object-centered coordinate system with labeled areas of interest as features, e.g., for faces this might consist of a vector for each fixation that indicates if the eyes, nose, or mouth is being fixated. An alternative is to use features that do not depend on the locations themselves and only encode relative motor activity, e.g., the direction and distance of each fixation from the previous fixation location. These representations could be used with both Fisher vector methods with relatively little effort.

Our definition of scanpath routines includes that they were acquired via implicit learning in the service of a task. We did not investigate how they are learned, but it is possible to create artificial tasks in the lab in which subjects learn efficient scanpaths without verbal instruction. For example, Rehder and Hoffman (2005) trained subjects to categorize stimuli consisting of three characters in a triangular array, where each character always appeared in the same location. Each character had two possible values, giving eight possible exemplars. Subjects were trained using feedback to categorize the eight exemplars into two equally-sized categories corresponding to one of the six category types of Shepard, Hovland and Jenkins (1961). Simple categories depended upon the value of one of the characters, while complex categories could require observing all three characters to classify. Over time, subjects developed efficient, stereotyped eye movements for the particular category type. Another example is the Desrochers et al.'s (2010) study. During each trial of their study, subjects (monkeys) were seated in front of an array of dots, with one randomly chosen dot providing a reward to the monkey when it was fixated. After several learning sessions, each monkey

acquired a stereotyped scanpath that visited all of the dots only once during each trial. This is more efficient than revisiting dots, but not all paths are equally efficient at minimizing the amount of time required. For both monkeys, this initial scanpath routine gradually became more efficient, but only one of the monkeys developed the optimal scanpath routine. Both Rehder and Hoffman (2005) and Desrochers et al. (2010) have been successfully modeled using reinforcement learning algorithms (Desrochers et al., 2010; Nelson and Cottrell, 2007). Similar mechanisms are thought to be implemented in the basal ganglia, which enables scanpath routines to be learned in humans and other animals (Hayhoe and Ballard, 2005; Hikosaka et al., 2000). The acquisition of efficient scanpath routines in more natural tasks, which lack artificial constraints, has yet to be studied.

Beyond inferring the task given to a person and their identity, our algorithms could be used to infer other traits. For example, these algorithms could potentially be used to diagnose Parkinson's disease or autism spectrum disorders, as long as a diagnostic stimuli could be identified to present to the subject. If this inference could be reliably made, MFPA would offer a low-cost diagnostic technique (especially since some eye trackers can now be purchased for less than $100). This approach has already shown success in predicting schizophrenia, attention deficit hyperactivity disorder, fetal alcohol spectrum disorder, and Parkinson's disease using algorithms similar to the summary statistics method (Benson et al., 2012; Tseng et al., 2013). Fisher vector methods that incorporate the spatio-temporal characteristics of scanpaths directly could lead to further improvements in disease diagnosis from eye movements.

5. Conclusions

In summary, we provide here the first direct evidence of scanpath routines for judging faces in humans. Consistent with other studies (Borji and Itti, 2014; Kanan et al., 2014), we found that a subject's task can be inferred solely from their eye movements. Our algorithms were most successful when judging observer's tasks from their own history of eye movement patterns, from which we conclude that, at least for these tasks, observers have idiosyncratic scanpath routines. It should be of considerable interest to further investigate whether experts in particular tasks converge on the same, optimal scanpath routines for their areas of expertise.

6. Acknowledgments

We thank the reviewers for their valuable comments, which substantially improved the manuscript. C.K. was affiliated with UC San Diego when this project was completed. This work was supported in part by NSF REU Site grant SMA-1005256 and NSF Science of Learning Center grants SBE-0542013 and SMA-1041755 to the Temporal Dynamics of Learning Center.

References

Althoff, R., Cohen, N., 1999. Eye movement-based memory. Journal of Experimental Psychology: Learning, Memory, & Cognition 25, 997–1010.

Benson, P., Beedie, S., Shephard, E., Giegling, I., Rujescu, D., St. Clair, D., 2012. Simple viewing tests can detect eye movement abnormalities that distinguish schizophrenia cases from controls with exceptional accuracy. Biological Psychiatry 71, 716–724.

Borji, A., Itti, L., 2014. Defending yarbus: Eye movements reveal observers' task. Journal of Vision 14 (3), 29.

Botev, Z., 2006. A novel nonparametric density estimator. Tech. rep., The University of Queensland.

Brainard, D. H., 1997. The psychophysics toolbox. Spatial Vision 10, 433–436.

Budesheim, T., DePaola, S., 1994. Beauty or the beast? the effects of appearance, personality, and issue information on evaluations of political candidates. Pers Soc Psych Bull 20, 339–348.

Chang, C., Lin, C., 2011. A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2 (3), 1–27.

Cornelissen, F., Peters, E., Palmer, J., 2002. The eyelink toolbox: Eye tracking with MATLAB and the psychophysics toolbox. Behavior Research Methods, Instruments & Computers, 34, 613–617.

Desrochers, T., Jin, D., Goodman, N., Graybiel, A., 2010. Optimal habits can develop spontaneously through sensitivity to local cost. Proc. of the National Academy of Sciences 107, 20512–20517.

Ebner, N., Riediger, M., Lindenberger, U., 2010. Faces–a database of facial expressions in young, middle-aged, and older women and men: Development and validation. Behavior Research Methods 42, 351–362.

Ekman, P., 1973. Darwin and facial expression: A century of research in review. Academic Press.

Everingham, M., Sivic, J., Zisserman, A., 2006. "Hello! My name is... Buffy" - Automatic naming of characters in TV video. In: Proc. of the British Machine Vision Conference.

Foulsham, T., Dewhurst, R., Nystrom, M., Jarodzka, H., Johansson, R., Underwood, G., Holmqvist, K., 2012. Comparing scanpaths during scene encoding and recognition: A multi-dimensional approach. Journal of Eye Movement Research 5, 1–14.

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

Greene, M. R., Liu, T., Wolfe, J. M., 2012. Reconsidering Yarbus: A failure to predict observer's task from eye movement patterns. Vision Research 62, 1–8.

Hayhoe, M., Ballard, D., 2005. Eye movements in natural behavior. Trends in Cognitive Sciences 9, 188–194.

Henderson, J., 2003. Human gaze control during real-world scene perception. Trends in Cognitive Science 7, 498–504.

Hikosaka, O., Takikawa, Y., Kawagoe, R., 2000. Role of the basal ganglia in the control of purposive saccadic eye movements. Physiological Reviews 80, 953–978.

Hsiao, J., Cottrell, G., 2008. Two fixations suffice in face recognition. Psychological Science 19 (10), 998–1006.

Izard, C. E., 1971. The Face of Emotion. Appleton-Century-Crofts.

Jaakkola, T., Diekhans, M., Haussler, D., 2000. A discriminative framework for detecting remote protein homologies. J. Comp. Biol. 7, 95–114.

Jaakkola, T., Haussler, D., 1998. Exploiting generative models in discriminative classifiers. In: Advances in Neural Information Processing Systems (NIPS-1998). pp. 487–493.

Joyce, C., 2000. Saving faces: Using eye movement, ERP, and SCR measures of face processing and recognition to investigate eyewitness identification. Ph.D. thesis, University of California San Diego.

Kanan, C., Ray, N., Bseiso, D., Hsiao, J., Cottrell, G., 2014. Predicting an observer's task using multi-fixation pattern analysis. In: 2014 Symposium on Eye Tracking Research and Applications (ETRA-2014).

Kleisner, K., Priplatova, L., Frost, P., Flegr, J., 2013. Trustworthy-looking face meets brown eyes. PLOS ONE 8, e53285.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D., Hawk, S., van Knippenberg, A., 2010. Presentation and validation of the Radboud faces database. Cognition & Emotion 24, 1377–1388.

Mehoudar, E., Arizpe, J., Baker, C., Yovel, G., 2014. Faces in the eye of the beholder: Unique and stable eye scanning patterns of individual observers. Journal of Vision 14 (7), 6.

Nelson, J., Cottrell, G., 2007. A probabilistic model of eye movements in concept formation. Neurocomputing 70, 2256–2272.

Nguyen, H., Isaacowitz, D., Rubin, P., 2009. Age- and fatigue-related markers of human faces: An eye tracking study. Opthamology 115, 355–360.

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

Oldfield, R., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113.

Perronnin, F., Sanchez, J., Mensink, T., 2010. Improving the Fisher kernel for large-scale image classification. In: European Conference on Computer Vision (ECCV-2010).

Peterson, M., Eckstein, M., 2012. Looking just below the eyes is optimal across face recognition tasks. Proc. of the National Academy of Sciences 109 (48), E3314–E3323.

Peterson, M., Eckstein, M., 2013. Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. Psychological Science 24, 1216–1225.

Rehder, B., Hoffman, A., 2005. Eyetracking and selective attention in category learning. Cognitive Psychology 51, 1–41.

Schyns, P., Bonnar, L., Gosselin, F., 2002. Show me the features! understanding recognition from the use of visual information. Psychological Science 13, 402–409.

Shepard, R., Hovland, C., Jenkins, H., 1961. Learning and memorization of classifications. Psychological Monographs 75, 1–42.

Simoyan, K., Parkhi, O., Vedaldi, A., Zisserman, A., 2013. Fisher vector faces in the wild. In: British Machine Vision Conference (BMVC-2013).

Spitz, H., Stark, L., Noton, D., 1971. Scanpaths and pattern recognition. Science 173, 753.

Sun, C., Nevatia, R., 2013. Active: Activity concept transitions in video event classification. In: IEEE International Conference on Computer Vision (ICCV-2013). pp. 913–920.

Sundelin, T., Lekander, M., Kecklund, G., Van Someren, E., Olsson, A., Axelsson, J., 2013. Cues of fatigue: effects of sleep deprivation on facial appearance. SLEEP 36 (9), 1355–1360.

Tatler, B., Wade, N., Kwan, H., Findlay, J., Velichkovsky, B., 2010. Yarbus, eye movements, and vision. i-Perception 1, 7–27.

Tseng, P., Cameron, I., Pari, G., Reynolds, J., Munoz, D., Itti, L., 2013. High-throughput classification of clinical populations from natural viewing eye movements. J. Neurol 260, 275–284.

van der Maaten, L., 2011. Learning discriminative Fisher kernels. In: Proc. 28th International Conference on Machine Learning (ICML-2011).

Vedaldi, A., Fulkerson, B., 2008. VLFeat: An open and portable library, `http://www.vlfeat.org/`.

Vinette, C., Gosselin, F., Schyns, P., 2004. Spatio-temporal dynamics of face recognition in a flash: its in the eyes. Cognitive Science 28, 289–301.

Wan, V., Renals, S., 2002. Evaluation of kernel methods for speaker verification and identification. In: Proc. IEEE ICASSP 2002. pp. 669–672.

Willis, J., Todorov, A., 2006. First impressions: making up your mind after a 100-ms exposure to a face. Psychological Science 17, 592–598.

Zaidel, D., Bava, S., Reis, V., 2003. Relationship between facial asymmetry and judging trustworthiness in faces. Laterality 8, 225–232.