Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2020) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on
Knowledge-Based and Intelligent Information & Engineering Systems

Discriminative Viewer Identification using Generative Models of Eye Gaze

Silvia Makowski[a,*], Lena A. Jäger[a], Lisa Schwetlick[b], Hans Trukenbrod[b], Ralf Engbert[b], Tobias Scheffer[a]

[a] *University of Potsdam, Department of Computer Science, 14482 Potsdam, Germany*
[b] *University of Potsdam, Department of Psychology, 14476 Potsdam, Germany*

Abstract

We study the problem of identifying viewers of arbitrary images based on their eye gaze. Psychological research has derived generative stochastic models of eye movements. In order to exploit this background knowledge within a discriminatively trained classification model, we derive Fisher kernels from different generative models of eye gaze. Experimentally, we find that the performance of the classifier strongly depends on the underlying generative model. Using an SVM with Fisher kernel improves the classification performance over the underlying generative model.

1. Introduction

Human eye movements are driven by a highly-complex interplay between voluntary and involuntary processes related to oculomotor control, high-level vision, cognition, and attention. While exploring a scene, the eyes move their focus three to four times per second on average by performing very fast movements, termed saccades [15]. This type of active perception is functional, since high visual acuity is only obtained within the fovea, a very small area on the retina. Visual uptake is limited to phases of relative gaze stability between the saccades, denoted as fixations [15]. The sequence of saccades and fixations that constitute the eye's response to a scene is referred to as *scanpath*. It has long been known that the way we move our eyes in response to a given stimulus is highly individual [28] and more recent psychological research has shown that these individual characteristics are reliable over time [4]. Hence, it has been proposed to use eye movements as a behavioral biometric characteristic [21, 5].

Psychologists have developed generative stochastic models in order to explain various aspects of scanpaths. The *SceneWalk* model [13] generates saccade amplitudes and directions of a viewer watching an image. A probabilistic

* Corresponding author.
E-mail address: silvia.makowski@uni-potsdam.de

model of reading [24] generates fixation durations, saccade amplitudes and durations, and the types of saccades (regressions to a previous word, refixations of the current word, skips ahead) that can occur during reading. Generative models constitute background knowledge about eye gaze, but they are optimized to maximize the likelihood of the observed scanpaths, rather than the accuracy of a discriminative task such as viewer identification. Fisher kernels allow the use of generative stochastic models as background knowledge to derive a feature representation from sequential data. For reading, an SVM with a Fisher kernel derived from a generative model of eye movements has been observed to performs substantially better at reader identification than the generative stochastic model itself [27]. This finding motivates our study on general scene viewing: starting from the *SceneWalk* model [13] and from a generative model for reading [24] which we adapt to general scene viewing and which we extend by incorporating additional features, we derive Fisher kernels that encode scanpaths in terms of their gradient for the generative stochastic models.

This paper is organized as follows. Section 2 defines the problem setting of viewer identification. Section 3 introduces two generative models of scanpaths, an adaptation of a reader identification model [24] and the SceneWalk model [13]. In Section 4, we develop the Fisher kernel function from these models. In Section 5, we evaluate our model and several baseline models. Section 6 concludes.

2. Problem Setting

When exploring a scene presented on a screen, a viewer generates a scanpath, which is a sequence $\mathbf{S} = ((q_1, d_1), \ldots, (q_T, d_T))$ of fixation positions q_t, measured in degrees of visual angle, and fixation durations d_t, measured in milliseconds. We study the problem of viewer identification and therefore train a model that selects the conjectured identity y of a viewer that generates a scanpath $\mathbf{S}$ on a certain picture, from a set of individuals that are known at training time. Training data consists of a set $\mathcal{D} = \{(\mathbf{S}_1, \mathbf{X}_1, y_1), ..., (\mathbf{S}_n, \mathbf{X}_n, y_n)\}$ of scanpaths $\mathbf{S}_1, ..., \mathbf{S}_n$ that have been obtained from subjects viewing pictures $\mathbf{X}_1, ..., \mathbf{X}_n$, labeled with viewers' identities $y_1, ..., y_n$.

3. Generative Models of Scanpaths

Let $p(\mathbf{S}|\mathbf{X}, \boldsymbol{\theta})$ be a parametric model of scanpaths given a picture $\mathbf{X}$. In a generative setting, viewer-specific models $p(\mathbf{S}|\mathbf{X}, \boldsymbol{\theta}_y)$ for user y can be estimated on viewer-specific data $\bar{\mathcal{D}}_y = \{(\mathbf{S}_i, \mathbf{X}_i)|(\mathbf{S}_i, \mathbf{X}_i, y_i) \in \mathcal{D}, y_i = y\}$ by maximum likelihood. At application time, the prediction for a scanpath $\mathbf{S}$ on a new picture $\mathbf{X}$ can be obtained as $y^* = \text{argmax}_y\, p(\mathbf{S}|\mathbf{X}, \boldsymbol{\theta}_y)$. For the discriminative setting we develop in Section 4, generative parameters are estimated on all training data $\bar{\mathcal{D}} = \{(\mathbf{S}_i, \mathbf{X}_i)|(\mathbf{S}_i, \mathbf{X}_i, y_i) \in \mathcal{D}\}$, and a Fisher score representation is derived from this generative model.

In this section, we modify a model for reader identification [24] to the case of viewer identification, and add velocity- and acceleration-based features to this model. We then review the *SceneWalk* model [13]. In Section 4, we will derive the Fisher kernel for both models and thus build a discriminative classifier for viewer identification.

3.1. Markov Model for Scene Viewing

In this section, we will review and adapt a model for reader identification [24] to reflect how viewers generate fixations while exploring a picture. The model assumes that the joint distribution over all fixation positions and durations is created by a Markov process:

$$p(q_1, \ldots, q_T, d_1, \ldots, d_T|\mathbf{X}, \boldsymbol{\theta}) = p(q_1, d_1|\mathbf{X}, \boldsymbol{\theta}) \prod_{t=1}^{T-1} p(q_{t+1}, d_{t+1}|q_t, \mathbf{X}, \boldsymbol{\theta}); \tag{1}$$

for this reason, we will refer to the model as *Markov model* in the following. To model the conditional distribution $p(q_t, d_t|q_{t-1}, \mathbf{X}, \boldsymbol{\theta})$ of the next fixation position q_t and duration d_t given the current fixation position q_t, the original model distinguishes between the *saccade types* of *regression* to a previous word, *refixation* of the current word before or after the current position, fixation of the *next word* or *skipping* one or more words. Our adaptation of the model

distinguishes four saccade types u: the scanpath can *maintain* the direction of the previous saccade up to ±45° ($u = 1$), change saccade direction to the *right* ($u = 2$), or to the *left* ($u = 3$) by more than 45°, or *reverse* direction by turning between 135° and 225° ($u = 4$). At time t, the model first draws a saccade type $u_t \sim p(u|\boldsymbol{\pi}) = \text{Mult}(u|\boldsymbol{\pi})$ from a multinomial distribution. Both the original and adapted models then draw a saccade amplitude $a_t \sim p(a|u_t)$, measured as the change of degrees of visual angle, from type-specific gamma distributions $p(a|u_t = u, \boldsymbol{\alpha}^a, \boldsymbol{\beta}^a) = \mathcal{G}(a|\alpha_u^a, \beta_u^a)$ for $u \in \{1, ..., 4\}$, where $\boldsymbol{\alpha}^a = \{\alpha_u^a | u \in \{1, ..., 4\}\}$, $\boldsymbol{\beta}^a = \{\beta_u^a | u \in \{1, ..., 4\}\}$ and $\mathcal{G}(\cdot|\alpha^a, \beta^a)$ is the gamma distribution parameterized by shape α^a and scale β^a. Analogously, the model draws a fixation duration $d_t \sim p(d|u_t, \boldsymbol{\alpha}^d, \boldsymbol{\beta}^d)$, also from type-specific gamma distributions $p(d|u_t = u, \boldsymbol{\alpha}^d, \boldsymbol{\beta}^d) = \mathcal{G}(d|\alpha_u^d, \beta_u^d)$ for $u \in \{1, ..., 4\}$, where $\boldsymbol{\alpha}^d = \{\alpha_u^d | u \in \{1, ..., 4\}\}$ and $\boldsymbol{\beta}^d = \{\beta_u^d | u \in \{1, ..., 4\}\}$.

3.1.1. Parameter Estimation

Given a set of k scanpaths on images $\bar{\mathcal{D}} = \{(\mathbf{S}_i, \mathbf{X}_i)\}$, all parameters are aggregated into a vector $\boldsymbol{\theta}$ and estimated by optimizing a maximum likelihood criterion $\boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{k} \ln p(\bar{\mathbf{S}}_i | \bar{\mathbf{X}}_i, \boldsymbol{\theta})$. Given $\bar{\mathcal{D}}$, all fixation positions q_t and saccade types u_t are known and the likelihood factorizes into separate likelihood terms depending on saccade type, amplitude, and duration parameters:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\pi}, \boldsymbol{\alpha}^a, \boldsymbol{\beta}^a, \boldsymbol{\alpha}^d, \boldsymbol{\beta}^d}{\text{argmax}} \left(\sum_{i=1}^{k} \sum_{t=1}^{T_i} \ln \text{Mult}(u_t^{(i)}|\boldsymbol{\pi}) + \sum_{i=1}^{k} \sum_{t=1}^{T_i} \ln p(a_t^{(i)}|u_t^{(i)}, \boldsymbol{\alpha}^a, \boldsymbol{\beta}^a) + \sum_{i=1}^{k} \sum_{t=1}^{T_i} \ln p(d_t^{(i)}|u_t^{(i)}, \boldsymbol{\alpha}^d, \boldsymbol{\beta}^d) \right). \quad (2)$$

3.2. Markov Model with Saccade Dynamics

Prior work on biometric identification using eye gaze has shown that saccade velocities, acceleration [16], and the relationship between peak velocity and amplitude of a saccade—referred to as *vigor*—convey information about viewer identity [30]. We therefore further extend the *Markov model* to include features that describe the saccade dynamics; we will study whether modeling these attributes of scanpaths contributes to identification accuracy. A survey of the data set used for evaluation in Section 5 shows that mean saccade velocities and accelerations follow Gamma distributions. Therefore, we extend the model to draw saccade mean velocities v as in Equation 3 and mean accelerations as in Equation 4 from type-specific gamma distributions:

$$v_t \sim p(v|u_t = u, \boldsymbol{\alpha}^v, \boldsymbol{\beta}^v) = \mathcal{G}(v|\alpha_u^v, \beta_u^v) \text{ for } u \in \{1, ..., 4\} \quad (3)$$

$$w_t \sim p(w|u_t = u, \boldsymbol{\alpha}^w, \boldsymbol{\beta}^w) = \mathcal{G}(w|\alpha_u^w, \beta_u^w) \text{ for } u \in \{1, ..., 4\}. \quad (4)$$

We define the peak acceleration-to-deceleration ratio of the horizontal saccade vector component r_t^x of saccade t as the ratio of the horizontal peak acceleration divided by the horizontal peak deceleration; the vertical peak acceleration-to-deceleration ratio r_t^y is defined in analogy; both ratios are governed by Gamma distributions:

$$r_t^x \sim p(r^x|u_t = u, \boldsymbol{\alpha}^{r^x}, \boldsymbol{\beta}^{r^x}) = \mathcal{G}(r^x|\alpha_u^{r^x}, \beta_u^{r^x}) \text{ for } u \in \{1, ..., 4\} \quad (5)$$

$$r_t^y \sim p(r^y|u_t = u, \boldsymbol{\alpha}^{r^y}, \boldsymbol{\beta}^{r^y}) = \mathcal{G}(r^y|\alpha_u^{r^y}, \beta_u^{r^y}) \text{ for } u \in \{1, ..., 4\}. \quad (6)$$

The relationship between the peak velocity v_t^{max}, amplitude a_t, and vigor g_t of a saccade t follows a parametric relationship $v_t^{max} = g_t \left(1 - e^{\frac{-a_t}{b}}\right)$ [3] with a global scalar rate parameter b. Following Rigas et al. [30], we fit rate parameter b in two steps. First, b and all g_t are jointly estimated via least-squares fitting on saccadic training data for each subject separately; then values b are averaged across subjects into a global rate parameter b^*. We incorporate the saccadic vigor for the vertical and horizontal components of each saccade and into the generative model via gamma

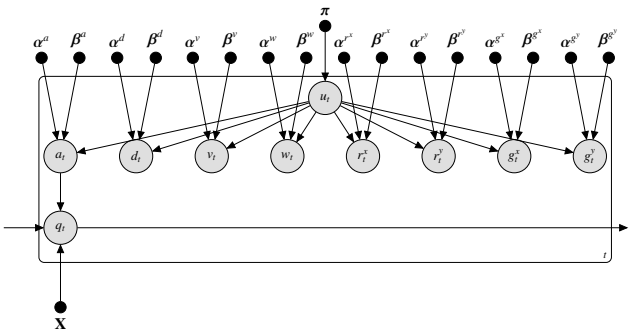

Fig. 1: Plate notation of the *Markov model with saccade dynamics*.

distributions:

$$g_t^x \sim p(g^x|u_t = u, \boldsymbol{\alpha}^{g^x}, \boldsymbol{\beta}^{g^x}) = \mathcal{G}(g^x|\alpha_u^{g^x}, \beta_u^{g^x}) \text{ for } u \in \{1, ..., 4\} \tag{7}$$

$$g_t^y \sim p(g^y|u_t = u, \boldsymbol{\alpha}^{g^y}, \boldsymbol{\beta}^{g^y}) = \mathcal{G}(g^y|\alpha_u^{g^y}, \beta_u^{g^y}) \text{ for } u \in \{1, ..., 4\}. \tag{8}$$

Figure 1 shows the *Markov model with saccade dynamics* as a plate diagram.

3.2.1. Parameter Estimation

Apart from the global rate parameter b, all model parameters are fitted for each user separately via maximum likelihood. The likelihood function is detailed in Appendix A.2 of the extended version of this paper [26].

3.3. The SceneWalk Model

SceneWalk [13] assumes the joint distribution over all fixation positions and durations of a scanpath to factorize as:

$$p(q_1, \ldots, q_T|\mathbf{X}, \boldsymbol{\theta}) = p(q_1|\mathbf{X}, \boldsymbol{\theta}) \prod_{t=1}^{T-1} p(q_{t+1}|q_1, ..., q_t, d_1, ..., d_t, \mathbf{X}, \boldsymbol{\theta}). \tag{9}$$

Here, $p(q_1|\mathbf{X}, \boldsymbol{\theta})$ is the likelihood of the first fixation position and can either be given by the experimental design (*e.g.,* by a fixation cross at a certain position that triggers the onset of an image) or the model itself [31].

When position $q_t = (i_t, j_t)$ is fixated, the model assigns a potential to each image pixel (i, j) to be the next saccade target q_{t+1}. This potential is obtained from an attentional component $\mathbf{A}_t$ and from an inhibitory component $\mathbf{F}_t$. Both components are based on Gaussian windows $\mathbf{G}_t^A$ and $\mathbf{G}_t^F$, respectively, centered at the position of q_t and with standard deviations σ_A and σ_F:

$$\mathbf{G}_t^{A/F}(i, j) = \frac{1}{2\pi\sigma_{A/F}^2} \exp\left(\frac{(i - i_t)^2 + (j - j_t)^2}{2\sigma_{A/F}^2}\right). \tag{10}$$

3.3.1. Attentional Component

The attentional component refers to the empirical *saliency* $\mathbf{H}$ of the image. The saliency map characterizes the intrinsic potential $\mathbf{H}(i, j)$ of image positions (i, j) to attract visual attention. The saliency is time-independent and can be obtained for each image separately, but globally across all viewers. It is common practice to estimate the saliency

by kernel density estimation with a bandwidth determined by Scott's Rule [2, 13]. The attentional component $\mathbf{A}_t$ is a dynamically evolving matrix that accesses the saliency matrix through a Gaussian window $\mathbf{G}^A_{q_t}$ which simulates the foveal area of high-acuity vision. The attentional component (Equation 11) changes over time at a rate of ω_A.

$$\mathbf{A}_t = \frac{\mathbf{G}^A_t \mathbf{H}}{\sum_{i,j} \mathbf{G}^A_t(i,j)\mathbf{H}(i,j)} + e^{-\omega_A d_t}\left(\mathbf{A}_{t-1} - \frac{\mathbf{G}^A_t \mathbf{H}}{\sum_{i,j} \mathbf{G}^A_t(i,j)\mathbf{H}(i,j)}\right) \tag{11}$$

3.3.2. Inhibitory Component

The inhibitory component $\mathbf{F}_t$ uses its Gaussian window to build up inhibition around the current fixation position and thus provokes an exploration of new regions of the image. It changes over time with a rate ω_F as in Equation 12.

$$\mathbf{F}_t = \frac{\mathbf{G}^F_t}{\sum_{i,j} \mathbf{G}^F_t(i,j)} + e^{-\omega_F d_t}\left(\mathbf{F}_{t-1} - \frac{\mathbf{G}^F_t}{\sum_{i,j} \mathbf{G}^F_t(i,j)}\right) \tag{12}$$

Both, the attentional and the inhibitory component, are calculated recursively, since they need the respective components of the previous fixation q_{t-1}.

3.3.3. Combined Potential for Target Selection

In the *SceneWalk* model, parameter c_F trades the attentional against the inhibitory component; λ and γ serve as regularization parameters. Equation 13 shows the resulting potential $\mathbf{U}_t$ for target selection.

$$\mathbf{U}_t = \frac{\mathbf{A}^\lambda_t}{\sum_{i,j} \mathbf{A}_t(i,j)^\lambda} - c_F \frac{\mathbf{F}^\gamma_t}{\sum_{i,j} \mathbf{F}_q(i,j)^\gamma} \tag{13}$$

3.3.4. Probabilities of image positions

Given a scanpath of fixation positions $q_1, ..., q_t$ and durations $d_1, ..., d_t$, the model calculates a probability for each possible image position to be the next fixation position q_{t+1} in the scanpath as a mixture of the normalized potential $\mathbf{U}_{q_t}$ and the uniform distribution over all image positions (i, j), weighted by a regularization parameter $\zeta \in [0, 1]$:

$$p(q_{t+1}|q_1, ..., q_t, d_1, ..., d_t, \theta) = (1-\zeta)\frac{\mathbf{U}_t(i_{t+1}, j_{t+1})}{\sum_{i,j} \mathbf{U}_t(i,j)} + \zeta \frac{1}{\sum_{i,j} 1}. \tag{14}$$

3.3.5. Parameter Estimation

In total, the parameter vector $\boldsymbol{\theta}$ of *SceneWalk* consists of eight parameters ω_A, ω_F, σ_A, σ_F, γ, λ, c_F, ζ. While [13] fit these parameters using maximum likelihood, we find this estimation technique to be numerically unstable and therefore resort to maximizing a regularized maximum likelihood criterion $\boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^k \ln p(\bar{\mathbf{S}}_i|\bar{\mathbf{X}}_i, \boldsymbol{\theta}) - \rho \sum_j \theta_j^2$.

4. Fisher Kernel

Fisher kernels [18] are a common framework that exploits generative probabilistic models as a representation of sequential or other structured instances within discriminative classifiers. The Fisher kernel approach projects structured input—here, scanpaths—into the gradient space of a generative probability model that was previously fitted to the training data via maximum likelihood. This section derives Fisher representations based on the generative models described in Sections 3.1, 3.2, and 3.3 to map scanpaths into feature vectors.

4.1. Fisher Kernel Function

The Fisher kernel function calculates the similarity of two scanpaths $\mathbf{S}_i$ and $\mathbf{S}_j$ as the inner product in the Riemannian manifold given by the class of probability models.

Definition 1 (Fisher kernel function). *Let $\boldsymbol{\theta}^*$ be the maximum likelihood estimate of a generative model on all training data. Let $\mathbf{S}_i$, $\mathbf{S}_j$ denote scanpaths on pictures $\mathbf{X}_i$, $\mathbf{X}_j$. The Fisher kernel between $\mathbf{S}_i$, $\mathbf{S}_j$ is $K((\mathbf{S}_i, \mathbf{X}_i), (\mathbf{S}_j, \mathbf{X}_j) = \mathbf{g}_i^\top \mathbf{I}^{-1} \mathbf{g}_j$ where $\mathbf{g}_i = \nabla_{\boldsymbol{\theta}} p(\mathbf{S}_i|\mathbf{X}_i, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ and where we employ the empirical version of the Fisher information matrix given by $\mathbf{I} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_i \mathbf{g}_i^\top$.*

The gradients of the log-likelihood functions of the Markov models are derived in Propositions 1 and 2.

4.2. Fisher Kernel for Markov Model

Proposition 1 (Gradient of log-likelihood of the Markov Model). *Let $\mathbf{S} = ((q_1, d_1), \ldots, (q_T, d_T))$ denote a scanpath obtained on an image $\mathbf{X}$. Let $a_1, ..., a_T$ denote the saccade amplitudes, and $u_1, ..., u_T$ denote the saccade types in $\mathbf{S}$. Define for $u \in \{1, 2, 3, 4\}$ the set $\{i_1^{(u)}, ..., i_{K_u}^{(u)}\} = \{i \in \{1, ..., T\} | u_i = u\}$. Let $\mathbf{a}_u = (|a_{i_1^{(u)}}|, ..., |a_{i_{K_u}^{(u)}}|)^\top$, $\mathbf{d}_u = (d_{i_1^{(u)}}, ..., d_{i_{K_u}^{(u)}})^\top$. Then the gradient of the logarithmic likelihood of the model defined in Section 3.1 is*

$$\mathbf{g} = \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{S}|\mathbf{X}, \boldsymbol{\theta}) = (\bar{\mathbf{g}}_1^\top, \bar{\mathbf{g}}_2^\top, \bar{\mathbf{g}}_3^\top, \bar{\mathbf{g}}_4^\top)^\top, \text{ where for } u \in \{1, 2, 3, 4\}: \quad \bar{\mathbf{g}}_u = \begin{pmatrix} \pi_u^{-1} K_u \\ \sum_{1 \le t \le T: u_t = u} \ln(a_t) - \psi(\alpha_u^a) - \beta_u^a \\ \frac{1}{\beta_u^a} \sum_{1 \le t \le T: u_t = u} \left(\frac{a_t}{\beta_u^a} - \alpha_u^a \right) \\ \sum_{1 \le t \le T: u_t = u} \ln(d_t) - \psi(\alpha_u^d) - \beta_u^d \\ \frac{1}{\beta_u^d} \sum_{1 \le t \le T: u_t = u} \left(\frac{d_t}{\beta_u^d} - \alpha_u^d \right) \end{pmatrix}.$$

A proof of Proposition 1 is given in Appendix A.1 of the extended version of this paper [26].

4.3. Fisher Kernel for Markov Model with Saccade Dynamics

Proposition 2 (Gradient of log-likelihood of the Markov Model with Saccade Dynamics). *In addition to Proposition 1 for the Markov Model for SceneViewing, let $v_1, ..., v_T$ denote the saccade mean velocities and $w_1, ..., w_T$ denote the saccade mean accelerations. Let $r_1^x, ..., r_T^x$ denote the horizontal and $r_1^y, ..., r_T^y$ the vertical peak-acceleration-to-deceleration ratio. Let $g_1^x, ..., g_T^x$ denote the horizontal and $g_1^y, ..., g_T^y$ the vertical saccade vigor in $\mathbf{S}$. Then the gradient of the logarithmic likelihood of the model defined in Appendix A.2 of the extended version of this paper [26] is*

$$\mathbf{g} = \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{S}|\mathbf{X}, \boldsymbol{\theta}) = (\bar{\mathbf{g}}_1^\top, \bar{\mathbf{g}}_2^\top, \bar{\mathbf{g}}_3^\top, \bar{\mathbf{g}}_4^\top)^\top, \text{ where for } u \in \{1, 2, 3, 4\}: \quad \bar{\mathbf{g}}_u = \begin{pmatrix} \pi_u^{-1} K_u \\ \sum_{1 \le t \le T: u_t = u} \ln(a_t) - \psi(\alpha_u^a) - \beta_u^a \\ \frac{1}{\beta_u^a} \sum_{1 \le t \le T: u_t = u} \left(\frac{a_t}{\beta_u^a} - \alpha_u^a \right) \\ \sum_{1 \le t \le T: u_t = u} \ln(d_t) - \psi(\alpha_u^d) - \beta_u^d \\ \frac{1}{\beta_u^d} \sum_{1 \le t \le T: u_t = u} \left(\frac{d_t}{\beta_u^d} - \alpha_u^d \right) \\ \sum_{1 \le t \le T: u_t = u} \ln(v_t) - \psi(\alpha_u^v) - \beta_u^v \\ \frac{1}{\beta_u^v} \sum_{1 \le t \le T: u_t = u} \left(\frac{v_t}{\beta_u^v} - \alpha_u^v \right) \\ \sum_{1 \le t \le T: u_t = u} \ln(w_t) - \psi(\alpha_u^w) - \beta_u^w \\ \frac{1}{\beta_u^w} \sum_{1 \le t \le T: u_t = u} \left(\frac{w_t}{\beta_u^w} - \alpha_u^w \right) \\ \sum_{1 \le t \le T: u_t = u} \ln(r_t^x) - \psi(\alpha_u^{r^x}) - \beta_u^{r^x} \\ \frac{1}{\beta_u^{r^x}} \sum_{1 \le t \le T: u_t = u} \left(\frac{r_t^x}{\beta_u^{r^x}} - \alpha_u^{r^x} \right) \\ \sum_{1 \le t \le T: u_t = u} \ln(r_t^y) - \psi(\alpha_u^{r^y}) - \beta_u^{r^y} \\ \frac{1}{\beta_u^{r^y}} \sum_{1 \le t \le T: u_t = u} \left(\frac{r_t^y}{\beta_u^{r^y}} - \alpha_u^{r^y} \right) \\ \sum_{1 \le t \le T: u_t = u} \ln(g_t^x) - \psi(\alpha_u^{g^x}) - \beta_u^{g^x} \\ \frac{1}{\beta_u^{g^x}} \sum_{1 \le t \le T: u_t = u} \left(\frac{g_t^x}{\beta_u^{g^x}} - \alpha_u^{g^x} \right) \\ \sum_{1 \le t \le T: u_t = u} \ln(g_t^y) - \psi(\alpha_u^{g^y}) - \beta_u^{g^y} \\ \frac{1}{\beta_u^{g^y}} \sum_{1 \le t \le T: u_t = u} \left(\frac{g_t^y}{\beta_u^{g^y}} - \alpha_u^{g^y} \right) \end{pmatrix}.$$

The likelihood of the newly-introduced features follows Gamma distributions in analogy to the distributions of durations and amplitudes in the *Markov model*; a proof of Proposition 2 is given in Appendix A.3 of the extended version of this paper [26].

4.4. Fisher Kernel for the SceneWalk Model

Given a scanpath $\mathbf{S} = ((q_1, d_1), \ldots, (q_T, d_T))$ obtained on an image $\mathbf{X}$, the gradient of the logarithmic likelihood under the SceneWalk model parameterized with $\boldsymbol{\theta} = (\zeta, c_F, \omega_A, \omega_F, \sigma_A, \sigma_F, \gamma, \lambda)$ is

$$\mathbf{g} = \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{S}|\mathbf{X}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ln p(q_1|\boldsymbol{\theta}, \mathbf{X}) + \sum_{t=2}^{T} \ln p(q_t|q_1, ..., q_{t-1}, d_1, ..., d_{t-1}, \boldsymbol{\theta}, \mathbf{X}).$$

We derive this gradient in Appendix A.4 of the extended version of this paper [26].

5. Empirical Study

This section explores the performance of the derived models and reference models for viewer identification.

5.1. Data Collection

We record the eye movements of 32 participants between the ages of 18 and 49 as they view a random subset of 106 images out of a set of 376 images. The images are colored photographs of natural scenes taken by the authors. Each photograph is presented once, for 8 seconds; the participants' task is to memorize the images. Participants sit at a viewing distance of 60 cm to the monitor, with their heads positioned in a chin rest. The monitor has a diagonal size of 61.4 cm, an aspect ratio of 16 by 10 (1920x1080 px), and a refresh rate of 100-120 Hz. The images are presented at a resolution of 1500×1000 px, and therefore subtend 48 degree by 28 degree of visual angle. We record participants' eye movements using an Eyelink 1000 video-based, desktop-mounted eye tracker with a sampling rate of 1000 Hz monocularly using the participant's dominant eye. All participants have normal or corrected-to-normal vision. From the raw samples recorded by the eyetracker, we extract the scanpaths using a velocity-based saccade detection algorithm [12].

5.2. Reference Methods

The natural reference methods for the *Fisher SVMs* are the underlying generative methods *Markov model*, *Markov model with saccade dynamics*, and *SceneWalk*. As an additional generative reference method, we use the model of [1] which has no Fisher kernel because it is nonparametric. Other prior work on biometric identification using eye movements varies with regard to the type of stimuli, and the features that are extracted from the scanpath. Stimuli used for viewer identification include viewing artificial stimuli [5, 20, 6, 19, 8, 35, 36, 33, 30], text documents [5, 17, 30], movies [22] or images [5, 8]. Most approaches are designed to identify viewers on a specific stimulus, for example by applying graph matching techniques to the scanpaths produced on a specific face image [29], or even by including a secondary identification task such as entering a PIN or password with the eye gaze [25, 23, 9, 10, 34, 7]. Approaches that can be applied to novel stimuli at test time extract different kinds of fixational and saccadic features, such as fixation durations [32, 14] or saccade amplitudes [14, 29, 19, 30], velocities [5, 32, 6, 29, 8, 19, 11, 30] and accelerations [29, 8, 11, 30], and either aggregate these over the whole scanpath [32, 17, 22, 8, 14], or compute the similarity of scanpaths by applying statistical tests to the distributions of the extracted features [16, 30]. As reference methods, we only consider methods that allow different stimuli for training and testing. As representative aggregational reference method, we choose the model by *Holland and Komogortsev (2011)*. As statistal reference approaches we use the seminal *CEM-B* method [16] and the current state-of-the-art model *CEM-B with saccade dynamics* [30].

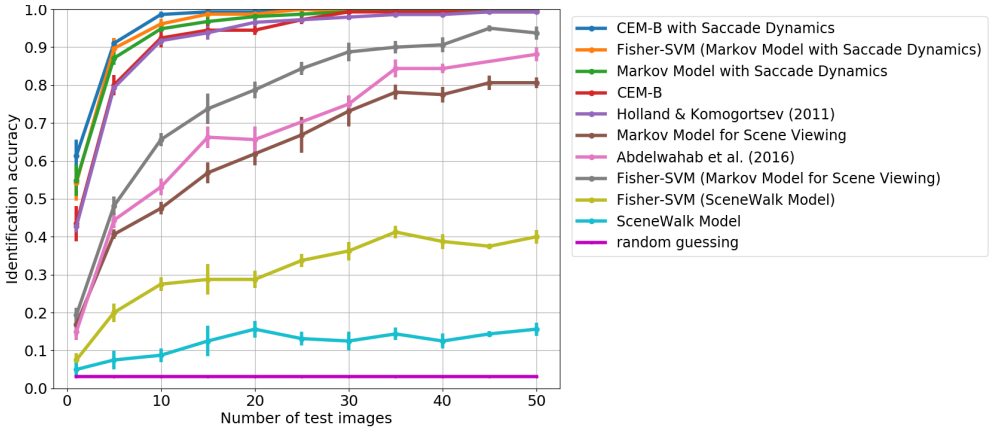

Fig. 2: Identification accuracy (32 subjects) of all compared models as a function of the number of images seen at test time. Error bars show the standard error. Training was performed with 50 images per subject.

5.3. Evaluation Setting

Each subject views a random subset of 106 photographs out of 376 photographs. We split the data into 50% training and 50% test data along photographs per subject. We average the identification accuracy across 5 random splits and study it as a function of the number of images seen at test time. All hyper-parameters of all methods are tuned by grid search using 3-fold cross validation on the training portion of the data.

5.4. Results

5.4.1. Identification Accuracy

Figure 2 compares the identification accuracy as a function of the number of images that have been viewed at test time. The *SceneWalk* model achieves the lowest identification accuracy; we attribute this to the fact that the classification can only be based on the saccade amplitudes and fixation durations since no other aspect of the scanpath is described by *SceneWalk*. The *Fisher SVM* for the *SceneWalk* model improves the classification accuracy dramatically ($p < 0.01$ for more than one test image). The *Markov model* has the second-lowest performance; in addition to saccade amplitudes and fixation durations it also models saccade durations and directions. Again, the *Fisher SVM* on the *Markov model* model improves the identification accuracy significantly over the generative model itself. The non-parametric model of *Abdelwahab et al. (2016)* outperforms the *Markov model* but is outperformed by the *Fisher SVM* based on the *Markov model*. The *Markov model with saccade dynamics* is the best-performing generative model. The performance comparison between this generative model and the *Fisher SVM* based on it is consistent with our previous observation that the *Fisher SVM* improves the classification accuracy of the underlying generative model; but in this case, the differences are not statistically significant. The *CEM-B with saccade dynamics* performs comparably to the *Fisher SVM* based on the *Markov model with saccade dynamics*; differences are not significant. *CEM-B with saccade dynamics* uses the largest feature set; in addition to the features extracted by the *Markov model with saccade dynamics*, it extracts the saccadic peak velocity, absolute starting times of fixations and saccades, and the fixation locations on the screen.

5.4.2. Execution Time of Identification

We compare the execution times for identifying a person based on input data from viewing one image and study them as a function of the number of persons in the training data. Figure 3 shows execution times on a single two-core CPU (Intel Core i7-6600U, 2.6GHz). The *Fisher SVM* (based on any generative model) is a generalized linear multi-class classifier; it has the lowest execution time, and the slope of the execution time over the number of persons (classes) is the lowest. The *CEM-B* method has a similar gradient but a higher absolute execution time. *CEM-B with saccade*

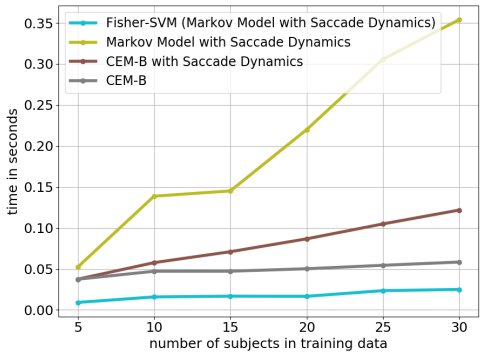

Fig. 3: Execution time in seconds to identify one subject, after viewing one single image, as a function of the number of subjects in training data.

dynamics extracts a larger set of distributional features and compares these features to the profiles of each user. The *Markov model with saccade dynamics* has to infer the likelihood of the observation sequence under each user-specific model and is therefore the slowest model by comparison.

6. Conclusion

We have adapted a generative model for eye gaze during reading [24] to scene viewing. We have integrated features that describe the saccade dynamics into this *Markov model with saccade dynamics*. Starting from these models and the known generative *SceneWalk* model for eye gaze during scene viewing, we have derived Fisher kernels for discriminative classification. Whereas generative models are trained to maximize the regularized likelihood of the observed gaze sequences, a *Fisher SVM* based on these generative models directly maximizes the classifier's ability to identify viewers based on their eye gaze. Experimentally, we find that the *Fisher SVM* generally improves identification accuracy compared to the underlying generative model. In terms of identification accuracy, the *Fisher SVM with saccade dynamics* performs comparably to *CEM-B with saccade dynamics* which extracts a larger set of distributional features from the scanpath; in terms of execution time, the *Fisher SVM* with any generative model of eye gaze is the fastest method in our comparison. We conclude that while Fisher SVMs improve the identification accuracy compared to the underlying generative model, the selection of features that are described by the generative model are crucial.

Acknowledgments

This work was partially funded by the German Science Foundation under grant SFB 1294 (project number 318763901).

References

[1] Abdelwahab, A., Kliegl, R., Landwehr, N., 2016. A semiparametric model for Bayesian reader identification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.
[2] Baddeley, A., Rubak, E., Turner, R., 2015. Spatial point patterns: methodology and applications with R. Chapman and Hall/CRC.
[3] Baloh, R.W., Sills, A.W., Kumley, W.E., Honrubia, V., 1975. Quantitative measurement of saccade amplitude, duration, and velocity. Neurology 25, 1065–1065.
[4] Bargary, G., Bosten, J.M., Goodbourn, P.T., Lawrance-Owen, A.J., Hogg, R.E., Mollon, J., 2017. Individual differences in human eye movements: An oculomotor signature? Vision Research 141, 157–169.
[5] Bednarik, R., Kinnunen, T., Mihaila, A., Fränti, P., 2005. Eye-movements as a biometric, in: Proceedings of the 14th Scandinavian Conference on Image Analysis (SCIA 2005), pp. 780–789.
[6] Cuong, N., Dinh, V., Ho, L.S.T., 2012. Mel-frequency cepstral coefficients for eye movement identification, in: 24th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 253–260.
[7] Cymek, D., Venjakob, A., Ruff, S., Lutz, O.M., Hofmann, S., Roetting, M., 2014. Entering PIN codes by smooth pursuit eye movements. Journal of Eye Movement Research 7, 1–11.

[8] Darwish, A., Pasquier, M., 2013. Biometric identification using the dynamic features of the eyes, in: 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6.
[9] De Luca, A., Weiss, R., Hußmann, H., An, X., 2007. Eyepass – eye-stroke authentication for public terminals, in: Extended Abstracts on Human Factors in Computing Systems (CHI EA '08), pp. 3003–3008.
[10] Dunphy, P., Fitch, A., Olivier, P., 2008. Gaze-contingent passwords at the ATM, in: 4th Conference on Communication by Gaze Interaction (COGAIN), pp. 59–62.
[11] Eberz, S., Rasmussen, K., Lenders, V., Martinovic, I., 2015. Preventing lunchtime attacks: Fighting insider threats with eye movement biometrics, in: Network and Distributed System Security (NDSS) Symposium.
[12] Engbert, R., Kliegl, R., 2003. Microsaccades uncover the orientation of covert attention. Vision Research 43, 1035–1045.
[13] Engbert, R., Trukenbrod, H.A., Barthelmé, S., Wichmann, F.A., 2015. Spatial statistics and attentional dynamics in scene viewing. Journal of Vision 15, 14–14.
[14] George, A., Routray, A., 2016. A score level fusion method for eye movement biometrics. Pattern Recognition Letters 82, 207–215.
[15] Henderson, J.M., Hollingworth, A., 1998. Eye movements during scene viewing: An overview, in: Eye guidance in reading and scene perception. Elsevier, pp. 269–293.
[16] Holland, C., Komogortsev, O., 2013. Complex eye movement pattern biometrics: Analyzing fixations and saccades, in: Proceedings of the International Conference on Biometrics.
[17] Holland, C., Komogortsev, O.V., 2011. Biometric identification via eye movement scanpaths in reading, in: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–8.
[18] Jaakkola, T., Haussler, D., 1999. Exploiting generative models in discriminative classifiers, in: Advances in neural information processing systems, pp. 487–493.
[19] Juhola, M., Zhang, Y., Rasku, J., 2013. Biometric verification of a subject through eye movements. Computers in Biology and Medicine 43, 42–50.
[20] Kasprowski, P., 2004. Human identification using eye movements. Ph.D. thesis. Silesian Unversity of Technology, Poland.
[21] Kasprowski, P., Ober, J., 2004. Eye movements in biometrics, in: International Workshop on Biometric Authentication, pp. 248–258.
[22] Kinnunen, T., Sedlak, F., Bednarik, R., 2010. Towards task-independent person authentication using eye movement signals, in: Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications (ETRA '10), pp. 187–190.
[23] Kumar, M., Garfinkel, T., Boneh, D., Winograd, T., 2007. Reducing shoulder-surfing by using gaze-based password entry, in: Proceedings of the 3rd Symposium on Usable Privacy and Security, pp. 13–19.
[24] Landwehr, N., Arzt, S., Scheffer, T., Kliegl, R., 2014. A model of individual differences in gaze control during reading, in: EMNLP, pp. 1810–1815.
[25] Maeder, A., Fookes, C., Sridharan, S., 2004. Gaze based user authentication for personal computer applications, in: Proceedings of the 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 727–730.
[26] Makowski, S., Jäger, L., Schwetlick, L., Trukenbrod, H., Engbert, R., Scheffer, T., 2020. Discriminative viewer identification using generative models of eye gaze. Technical Report. arXiv:2003.11399.
[27] Makowski, S., Jäger, L.A., Abdelwahab, A., Landwehr, N., Scheffer, T., 2018. A discriminative model for identifying readers and assessing text comprehension from eye movements, in: Proceedings of the European Conference on Machine Learning (ECML).
[28] Noton, D., Stark, L., 1971. Scanpaths in eye movements during pattern perception. Science 171, 308–311.
[29] Rigas, I., Economou, G., Fotopoulos, S., 2012. Biometric identification based on the eye movements and graph matching techniques. Pattern Recogntion Letters 33, 786–792.
[30] Rigas, I., Komogortsev, O., Shadmehr, R., 2016. Biometric recognition via eye movements: Saccadic vigor and acceleration cues. ACM Transactions on Applied Perception 13, 6.
[31] Schütt, H.H., Rothkegel, L.O., Trukenbrod, H.A., Reich, S., Wichmann, F.A., Engbert, R., 2017. Likelihood-based parameter estimation and comparison of dynamical cognitive models. Psychological review 124, 505.
[32] Silver, D.L., Biggs, A., 2006. Keystroke and eye-tracking biometrics for user identification, in: Proceedings of the 2006 International Conference on Artificial Intelligence (ICAI 2006), pp. 344–348.
[33] Srivastava, N., Agrawal, U., Roy, S., Tiwary, U.S., 2015. Human identification using linear multiclass svm and eye movement biometrics, in: 8th International Conference on Contemporary Computing (IC3), pp. 365–369.
[34] Weaver, J., Mock, K., Hoanca, B., 2011. Gaze-based password authentication through automatic clustering of gaze points, in: 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2749–2754.
[35] Yoon, H.J., Carmichael, T.R., Tourassi, G., 2014. Gaze as a biometric, in: Proceedings of the 2014 SPIE Medical Imaging Conference: Image Perception, Observer Performance, and Technology Assessment.
[36] Zhang, Y., Laurikkala, J., Juhola, M., 2014. Biometric verification of a subject with eye movements, with special reference to temporal variability in saccades between a subject's measurements. International Journal of Biometrics 6, 75–94.