Probabilistic, Deep, and Metric Learning for Biometric Identification from Eye Movements

Kumulative Dissertation

zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin Informatik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
Institut für Informatik
der Universität Potsdam

von
Ahmed Abdelwahab

Potsdam, den 15 November 2019

Published online in the
Institutional Repository of the University of Potsdam:
https://doi.org/10.25932/publishup-46798
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-467980

Abstract

A central insight from psychological studies on human eye movements is that eye movement patterns are highly individually characteristic. They can, therefore, be used as a biometric feature, that is, subjects can be identified based on their eye movements. This thesis introduces new machine learning methods to identify subjects based on their eye movements while viewing arbitrary content. The thesis focuses on probabilistic modeling of the problem, which has yielded the best results in the most recent literature. The thesis studies the problem in three phases by proposing a purely probabilistic, probabilistic deep learning, and probabilistic deep metric learning approach.

In the first phase, the thesis studies models that rely on psychological concepts about eye movements. Recent literature illustrates that individual-specific distributions of gaze patterns can be used to accurately identify individuals. In these studies, models were based on a simple parametric family of distributions. Such simple parametric models can be robustly estimated from sparse data, but have limited flexibility to capture the differences between individuals. Therefore, this thesis proposes a semiparametric model of gaze patterns that is flexible yet robust for individual identification. These patterns can be understood as domain knowledge derived from psychological literature. Fixations and saccades are examples of simple gaze patterns. The proposed semiparametric densities are drawn under a Gaussian process prior centered at a simple parametric distribution. Thus, the model will stay close to the parametric class of densities if little data is available, but it can also deviate from this class if enough data is available, increasing the flexibility of the model. The proposed method is evaluated on a large-scale dataset, showing significant improvements over the state-of-the-art.

Later, the thesis replaces the model based on gaze patterns derived from psychological concepts with a deep neural network that can learn more informative and complex patterns from raw eye movement data. As previous work has shown that the distribution of these patterns across a sequence is informative, a novel statistical aggregation layer called the quantile layer is introduced. It explicitly fits the distribution of deep patterns learned directly from the raw eye movement data. The proposed deep learning approach is end-to-end learnable, such that the deep model learns to extract informative, short local patterns while the quantile layer learns to approximate the distributions of these patterns. Quantile layers are a generic approach that can converge to standard pooling layers or have a more detailed description of the features being pooled, depending on the problem. The proposed model is evaluated in a large-scale study using the eye movements of subjects viewing arbitrary visual input. The model improves upon the standard pooling layers and other statistical aggregation layers proposed in the literature. It also improves upon the state-of-the-art eye movement biometrics by a wide margin.

Finally, for the model to identify any subject — not just the set of subjects it is trained on — a metric learning approach is developed. Metric learning learns a distance function over instances. The metric learning model maps the instances into a metric space, where sequences of the same individual are close, and sequences of different individuals are further apart. This thesis introduces a deep metric learning approach with distributional embeddings. The approach represents sequences as a set of continuous distributions in a metric space; to achieve this, a new loss function based on Wasserstein distances is introduced. The proposed method is evaluated on multiple domains besides eye movement biometrics. This approach outperforms the state of the art in deep metric learning in several domains while also outperforming the state of the art in eye movement biometrics.

Contents

Chapter 1

Introduction

In the information age, human identification is central to accessing information. Due to the increased availability of cheap hardware and sensors, new methods of identification arise. Biometric studies have gained attraction as a result, enabling authentication without the need to possess an identification card or memorize a password; thus, it cannot be misplaced, lost, or forgotten. Biometrics are measures related to human characteristics, either physical or behavioral. This thesis is concerned with behavioral biometrics embodied in eye movements. In our daily life, we are actively scanning the world around us through our eye movements at conscious and unconscious levels. Eye movements are studied widely in psychology, cognitive science, and applied research fields, which reveal that eye movements are strongly correlated with cognitive and perceptive processes [Rayner 98, Duchowski 02]. Several psychological studies have demonstrated that human eye movements are individually characteristic during various tasks [Afflerbach 15, Dixon 51, Huey 08, Poynter 13, Rayner 07]. These individual differences in eye movements qualify eye movements to act as a biometric measure. Eye movements can be used for unobtrusive identification [Landwehr 14, Kinnunen 10]. The advantage of unobtrusive identification is in eliminating the need for challenge protocols or extra effort by users for authentication, enabling systems to continuously identify or verify users. This thesis uses machine learning for identifying humans from their eye movements while viewing arbitrary content.

The psychological literature has studied typical short, local patterns that arise in human eye movements. From the point of view of this thesis, such patterns can be seen as domain knowledge derived from eye movement research. Recent work has shown that the distribution of these patterns within eye movements is informative for subject identification [Landwehr 14]. One focus of this thesis is thus on machine learning models that are probabilistic, in the sense that they capture such distributions. The thesis exploration of the problem starts with the development of a probabilistic model that is based on psychological domain knowledge in the form of short local patterns such as saccades and fixations. As eye movement data is low-level sensor data, deep neural networks might be able to learn more complex and informative patterns directly from the raw eye movement data. The thesis, therefore, next investigates deep neural networks for biometric identification from eye movements. The thesis specifically introduces a novel statistical aggregation layer for deep neural networks, which can fit the distribution of learned deep patterns within a sequence, thereby integrating probabilistic modeling and deep learning. As in most applica-

tions of biometric identification, there is no fixed set of classes (subjects), the thesis finally investigates metric learning approaches for eye movement biometrics. The thesis proposes a novel approach that integrates deep metric learning with probabilistic modeling.

1.1 Eye Movements

Humans are not simply exposed to an incoming flow of visual data, they actively select their visual input through their eye movements. Eye movements have shown to be the result of visual attention, cognition, and motor control [Liversedge 00, Henderson 03, Kliegl 06], holding significant information about identities. Eye movements can be accurately measured by devices called eye trackers. Eye trackers emit infrared light that is reflected by the eye. Eye trackers track the reflection relative to the position of the pupil, and thereby determine the direction of the gaze. The position the subject looks at can then be calculated by projecting the gaze direction on the screen at a constant temporal resolution. Thus, the raw gaze movement data is a series of coordinates sampled at a fixed rate depending on the capability and settings of the eye tracker device. As such, an individual viewing an input on a screen results in a time-series of coordinates.

According to the psychological literature, every gaze movement sequence can be expressed as a number of *fixations* that last between 200–300 milliseconds on average, and fast movements between these fixations called *saccades* [Rayner 98, Salvucci 00]. A fixation is when the eye is fairly still so that it can take in the information at the fixated point. A fixation is made up of multiple gaze points in the raw sequence, usually described by the fixation duration. A saccade is the movement between fixations or points of interest and lasts between 20–40 milliseconds, on average. A saccade amplitude is the distance skipped by the eye, which is the distance between two fixations. The duration of a saccade and its amplitude are linearly correlated. Saccades are usually described by their amplitudes. Alternating between fixations and saccades allows our brain to have an efficient perception by skipping over areas that have no information and fixating on areas that interest us. Studies dealing with eye movements usually preprocess the raw time-series to sequences of fixations and saccades, such that fixations are described by their durations and fixating point coordinates and saccades are described by their amplitudes. Different people have different fixation durations and different saccade amplitudes. For example, while reading, some tend to have many fixations, but others tend to skip through. As another example, when watching a video clip, some subjects tend to have longer fixations with smooth pursuit, while others tend to have larger saccade amplitudes.

1.2 Problem Setup

Eye movement biometrics can be formalized as a sequence classification problem. Considering each individual as a class, the task is to classify a sequence of gaze movements. The sequence is generated by an individual while viewing an arbitrary input on the screen. This thesis is concerned with a setting where the input is arbitrary in contrast to studies that have investigated the problem for fixed inputs or designed stimuli. The dataset is a collection of sequences generated while viewing arbitrary

inputs, such that each sequence is labeled by a class (individual). To simulate an application setting in which subjects are viewing arbitrary inputs, all available data should be split into training and test data in such a way that the test data contains visual inputs not seen by the model in the training data. The training data contains eye movement recordings for the subjects. It is used to train the models to distinguish between sequences of different subjects. During testing, sequences generated while viewing test inputs by unknown subjects are used to test the abilities of models to predict the identities of those subjects.

Modeling the biometric problem as a classification problem would require the classification model to be retrained every time a new user joins the system. For a more general approach, the problem can be formalized as a metric learning problem, in which the model learns a generic distance function to compare instances. In metric learning, distances between sequences generated by the same subject are minimized while distances between sequences generated by different subjects are maximized. Metric learning summarizes the whole sequence with an embedding, which represents individual characteristics in a low dimensional space. As in the classification setting, the test data contains eye movement sequences obtained on inputs not seen in the training data. Additionally, data should be split such that the test data contains sequences generated by novel subjects. Thus, the test dataset has a set of subjects that the model did not see during training, testing the ability of the model to generalize beyond the classes in the training data.

Two tasks are usually considered in biometric studies: multi-class classification (identification) and binary classification (verification). In multi-class classification, the model identifies to whom a given eye movement sequence belongs, out of all the known subjects. As the number of known subjects increases, identification gets more difficult, which is why identification performance is measured as a function of the number of subjects that the model needs to distinguish. In contrast, in binary classification, the model has to verify whether a given sequence of eye movements is generated by a given individual. The verification setting is independent of the number of classes and can deal with instances that do not belong to any of the classes present during training (imposters). The multi-class classification setting can directly detect imposters only in a metric learning problem setting.

This thesis investigates the stated problems in three incremental studies, which the following sections introduce. The thesis is initially inspired by psychological studies, which segment the raw eye movement data into sequences of several types of saccades and fixations. Based on these sequences, the thesis introduces the probabilistic classification model in Section 1.3 for reader identification. In Section 1.4, the thesis abandons the psychological concepts and introduces end-to-end learnable deep neural networks that classify the raw eye movement time-series data. In this section, the thesis introduces a novel method for fitting distributions of learned patterns within deep neural networks. The thesis then continues with metric learning in Section 1.5. In the section, the thesis introduces a generic probabilistic deep metric learning approach, in which an instance is represented by a set of distributions. The proposed method is evaluated on several biometric domains including eye movement biometrics.

1.3 A Semiparametric Probabilistic Model of Eye Movements

The thesis begins by developing a probabilistic model based on psychological concepts about eye movements, which have emphasized preprocessing eye movement sequences into saccades and fixations. In psychological studies, the gaze movements during reading are divided into four different types: refixating the current word (*refixation*), fixating the next word (*next word movement*), moving the fixation to a word after the next word (*forward skip*), and regressing to fixate on a word occurring earlier than the currently fixated word (*regression*) [Heister 12]. The latest state-of-the-art study [Landwehr 14] in reader identification at the time the thesis was started had illustrated that modeling reader-specific distributions over saccade amplitudes and fixation durations of these types can help identify readers accurately. The study approximated these distributions with simple parametric Gamma distributions. By approximating the true distribution with a simple parametric density, the model can be robust to sparse data, but it might not be flexible enough to fit the differences between distributions of different readers. Instead of employing simple parametric modeling for the distribution, the study in Chapter 2 develops a semiparametric model that allows the fitted distribution to be more flexible. The densities are inferred under a Gaussian process prior, centered at the parametric family. If the data is sparse, the posterior will favor the simple parametric family, reducing the flexibility of the model and minimizing overfitting. If more data is available, the model will deviate from the simple parametric prior towards more general densities. By adjusting the kernel function for the Gaussian process prior, any density function can be represented. However, the inference process is nontrivial as the text structure induces truncations to the semiparametric distributions, which is different for each observed sample. The thesis introduces a Metropolis-Hastings based algorithm for Bayesian inference that can reflect these observation-specific truncations. The proposed approach can fit the individual-specific distribution, balancing between robustness and flexibility. In a large-scale study, it demonstrates much better accuracy than the state-of-the-art methods.

In this work [Abdelwahab 16], and jointly with Niels Landwehr, I developed the mathematical framework, probabilistic model, and algorithms for inference and learning. I implemented the method and baselines, and designed and implemented all the empirical studies. Moreover, I contributed to the writing of the manuscript. In another study [Makowski 18], where my contributions are limited and therefore not part of this thesis, I adapted the semiparametric model and the fully parametric model [Landwehr 14] to work on the free viewing of a full document to assess the text comprehension of the readers and to identify readers.

1.4 Statistical Aggregation in Deep Neural Networks

Gaze movements are low-level sensor data, for which deep learning excels. The next study thus abandons the psychological concepts in favor of deep learning, which can automatically learn informative patterns from raw data. By examining the previous

model introduced in Section 1.3 and other state-of-the-art models from the literature, it is clear that distributions of short-term local patterns (saccades and fixations) are very informative for identifying individuals from their eye movements. Therefore, the study in Chapter 3 proposes an end-to-end learnable neural network architecture that extracts informative local patterns and characterizes their distributions. The study introduces a parametrized learnable statistical aggregation layer called a *quantile layer*. The quantile layer allows the network to explicitly fit the distributions of learned deep patterns and allows the network to process variable-length sequences.

Specifically, a convolution architecture is used to learn local, short-term patterns in a sequence, while the quantile layer is used to describe the distribution of the learned patterns. The quantile layer approximates the quantile function (the inverse of the cumulative distribution function) of the filter activations across the entire sequence and samples the function at multiple learnable points, which may differ from one filter to another. The whole method is end-to-end learnable directly from raw time-series data. The quantile layer generalizes standard pooling layers and can converge to maximum pooling or average pooling (for zero skewness distribution) but can also be more expressive. The empirical study shows that the proposed deep learning method outperforms the state of the art in eye movement biometrics by a large margin.

In this work [Abdelwahab 19b], I designed and implemented the proposed method as well as all baseline methods. Furthermore, I designed and conducted the empirical study, including the data preprocessing and adaptation required by the different baseline methods. I contributed to the writing of the manuscript.

1.5 Deep Metric Learning with Distributional Embeddings

One of the weaknesses of the previous study introduced in Section 1.4 is that the model has to be retrained every time a new class is added. Therefore, every time a new subject joins or enrolls, there must be enough training data for the subject, and the model must be retrained. The model also cannot directly detect imposters because it can only classify a given sequence into a fixed set of classes, namely the subjects present in the training data. In Chapter 4, the thesis does not deal with the biometric problem as a sequence classification problem but as a sequence metric learning problem. The chapter extends the previous approach introduced in Section 1.4 to describe an instance in a metric space by a set of continuous distributions, using an interpolated version of the quantile layer. Unlike existing deep metric learning methods, which use a fixed-point vector for representation, the chapter describes a sequence by a set of distributions. The distance between the embeddings needs to reflect the fact that the embeddings are distributions. This leaves the current deep metric learning losses — based on fixed-point vectors — not applicable. The study, therefore, introduces a loss based on Wasserstein distances, which satisfies the metric properties. Compared to other statistical distance functions — such as Kulback-Leibler or Jensen-Shannon divergence — the advantage of using Wasserstein distances is that it considers the space on which the random variable is defined. This property enables Wasserstein distances to be continuous and avoid the zero-gradient problem observed in other distances between distributions. This enables end-to-end

learning of distributional embeddings by directly optimizing model parameters according to a metric learning loss based on Wasserstein distances. The empirical study in this work includes multiple domains and is not limited to eye movement biometrics, as the proposed method is a general method to learn sequence embeddings. In all studied domains in the empirical study, the proposed loss function outperforms the state-of-the-art loss function in deep metric learning. Furthermore, the empirical study demonstrates that for eye movement biometrics the proposed approach has a better performance than the previous model introduced in Section 1.4.

In this work [Abdelwahab 19a], I developed the idea, mathematical framework, architecture, and algorithmic details of the method. I also implemented the method and all the baseline methods, and I designed and conducted all the experiments. I wrote the manuscript jointly with Niels Landwehr.

A Semiparametric Model for Bayesian Reader Identification

Ahmed Abdelwahab[1] and **Reinhold Kliegl**[2] and **Niels Landwehr**[1]
[1] Department of Computer Science, Universität Potsdam
August-Bebel-Straße 89, 14482 Potsdam, Germany
{abdelwahab,landwehr}@cs.uni-potsdam.de
[2] Department of Psychology, Universität Potsdam
Karl-Liebknecht-Straße 24/25, 14476 Potsdam OT/Golm
kliegl@uni-potsdam.de

Abstract

We study the problem of identifying individuals based on their characteristic gaze patterns during reading of arbitrary text. The motivation for this problem is an unobtrusive biometric setting in which a user is observed during access to a document, but no specific challenge protocol requiring the user's time and attention is carried out. Existing models of individual differences in gaze control during reading are either based on simple aggregate features of eye movements, or rely on parametric density models to describe, for instance, saccade amplitudes or word fixation durations. We develop flexible semiparametric models of eye movements during reading in which densities are inferred under a Gaussian process prior centered at a parametric distribution family that is expected to approximate the true distribution well. An empirical study on reading data from 251 individuals shows significant improvements over the state of the art.

1 Introduction

Eye-movement patterns during skilled reading consist of brief fixations of individual words in a text that are interleaved with quick eye movements called *saccades* that change the point of fixation to another word. Eye movements are driven both by low-level visual cues and high-level linguistic and cognitive processes related to text understanding; as a reflection of the interplay between vision, cognition, and motor control during reading they are frequently studied in cognitive psychology (Kliegl et al., 2006; Rayner, 1998). Computational models (Engbert et al., 2005; Reichle et al., 1998) as well as models based on machine learning (Matties and Søgaard, 2013; Hara et al., 2012) have been developed to study how gaze patterns arise based on text content and structure, facilitating the understanding of human reading processes.

A central observation in these and earlier psychological studies (Huey, 1908; Dixon, 1951) is that eye movement patterns strongly differ between individuals. Holland et al. (2012) and Landwehr et al. (2014) have developed models of individual differences in eye movement patterns during reading, and studied these models in a biometric problem setting where an individual has to be identified based on observing her eye movement patterns while reading arbitrary text. Using eye movements during reading as a biometric feature has the advantage that it suffices to observe a user during a routine access to a device or document, without requiring the user to react to a specific challenge protocol. If the observed eye movement sequence is unlikely to be generated by an authorized individual, access can be terminated or an additional verification requested. This is in contrast to approaches where biometric identification is based on eye movements in response to an artificial visual stimulus, for example a moving (Kasprowski and Ober, 2004; Komogortsev et al., 2010; Rigas et al., 2012b; Zhang and Juhola, 2012) or fixed (Bednarik et al., 2005) dot on a computer screen, or a specific image stimulus (Rigas et al., 2012a).

The model studied by Holland & Komogortsev (2012) uses aggregate features (such as average fixation duration) of the observed eye movements. Landwehr et al. (2014) showed that readers can be identified more accurately with a model that captures aspects of individual-specific distributions over

eye movements, such as the distribution over fixation durations or saccade amplitudes for word refixations, regressions, or next-word movements. Some of these distributions need to be estimated from very few observations; a key challenge is thus to design models that are flexible enough to capture characteristic differences between readers yet robust to sparse data. Landwehr et al. (2014) used a fully parametric approach where all densities are assumed to be in the gamma family; gamma distributions were shown to approximate the true distribution of interest well for most cases (see Figure 1). This model is robust to sparse data, but might not be flexible enough to capture all differences between readers.

The model we study in this paper follows ideas developed by Landwehr et al. (2014), but employs more flexible semiparametric density models. Specifically, we place a Gaussian process prior over densities that concentrates probability mass on densities that are close to the gamma family. Given data, a posterior distribution over densities is derived. If data is sparse, the posterior will still be sharply peaked around distributions in the gamma family, reducing the effective capacity of the model and minimizing overfitting. However, given enough evidence in the data, the model will also deviate from the gamma-centered prior—depending on the kernel function chosen for the GP prior, any density function can in principle be represented. Integrating over the space of densities weighted by the posterior yields a marginal likelihood for novel observations from which predictions are inferred. We empirically study this model in the same setting as studied by Landwehr et al. (2014), but using an order of magnitude more individuals. Identification error is reduced by more than a factor of three compared to the state of the art.

The rest of the paper is organized as follows. After defining the problem setting in Section 2, Section 3 presents the semiparametric probabilistic model. Section 4 discusses inference, Section 5 presents an empirical study on reader identification.

2 Problem Setting

Assume R different readers, indexed by $r \in \{1, \ldots, R\}$, and let $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ denote a set of texts. Each $r \in \mathcal{R}$ generates a set of eye movement patterns $\mathcal{S}^{(r)} = \{\mathbf{S}_1^{(r)}, \ldots, \mathbf{S}_n^{(r)}\}$ on $\mathcal{X}$, by

$$\mathbf{S}_i^{(r)} \sim p(\mathbf{S}|\mathbf{X}_i, r, \mathbf{\Gamma})$$

where $p(\mathbf{S}|\mathbf{X}_i, r, \mathbf{\Gamma})$ is a reader-specific distribution over eye movement patterns given a text $\mathbf{X}_i$. Here, r is a variable indicating the reader generating the sequence, and $\mathbf{\Gamma}$ is a true but unknown model that defines all reader-specific distributions. We assume that $\mathbf{\Gamma}$ can be broken down into reader-specific models, $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k)$, such that the distribution

$$p(\mathbf{S}|\mathbf{X}_i, r, \mathbf{\Gamma}) = p(\mathbf{S}|\mathbf{X}_i, \boldsymbol{\gamma}_r) \tag{1}$$

is defined by the partial model $\boldsymbol{\gamma}_r$. We aggregate the observations of all readers on the training data into a variable $\mathcal{S}^{(1:R)} = (\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(R)})$.

We follow a Bayesian approach, defining a prior $p(\mathbf{\Gamma})$ over the joint model that factorizes into priors over reader-specific models, $p(\mathbf{\Gamma}) = \prod_{r=1}^{R} p(\boldsymbol{\gamma}_r)$. At test time, we observe novel eye movement patterns $\bar{\mathcal{S}} = \{\bar{\mathbf{S}}_1, \ldots, \bar{\mathbf{S}}_m\}$ on a novel set of texts $\bar{\mathcal{X}} = \{\bar{\mathbf{X}}_1, \ldots, \bar{\mathbf{X}}_m\}$ generated by an unknown reader $r \in \mathcal{R}$. We assume a uniform prior over readers, that is, each $r \in \mathcal{R}$ is equally likely to be observed at test time. The goal is to infer the most likely reader to have generated the novel eye movement patterns. In a Bayesian setting, this means inferring the most likely reader given the training observations $(\mathcal{X}, \mathcal{S}^{(1:R)})$ and test observation $(\bar{\mathcal{X}}, \bar{\mathcal{S}})$:

$$r_* = \arg\max_{r \in \mathcal{R}} p(r|\bar{\mathcal{X}}, \bar{\mathcal{S}}, \mathcal{X}, \mathcal{S}^{(1:R)}). \tag{2}$$

We can rewrite Equation 2 to

$$\begin{aligned} r_* &= \arg\max_{r \in \mathcal{R}} p(\bar{\mathcal{S}}|r, \bar{\mathcal{X}}, \mathcal{X}, \mathcal{S}^{(1:R)}) && (3) \\ &= \arg\max_{r \in \mathcal{R}} \int p(\bar{\mathcal{S}}|r, \bar{\mathcal{X}}, \mathbf{\Gamma}) p(\mathbf{\Gamma}|\mathcal{X}, \mathcal{S}^{(1:R)}) \mathrm{d}\mathbf{\Gamma} \\ &= \arg\max_{r \in \mathcal{R}} \int p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \boldsymbol{\gamma}_r) p(\boldsymbol{\gamma}_r|\mathcal{X}, \mathcal{S}^{(r)}) \mathrm{d}\boldsymbol{\gamma}_r && (4) \end{aligned}$$

where

$$p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \boldsymbol{\gamma}_r) = \prod_{i=1}^{m} p(\bar{\mathbf{S}}_i|\bar{\mathbf{X}}_i, \boldsymbol{\gamma}_r) \tag{5}$$

$$p(\boldsymbol{\gamma}_r|\mathcal{X}, \mathcal{S}^{(r)}) \propto p(\boldsymbol{\gamma}_r) \prod_{i=1}^{n} p(\mathbf{S}_i^{(r)}|\mathbf{X}_i, \boldsymbol{\gamma}_r). \tag{6}$$

In Equation 3 we exploit that readers are uniformly chosen at test time, and in Equation 4 we exploit the factorization $p(\boldsymbol{\Gamma}) = \prod_{r=1}^{R} p(\boldsymbol{\gamma}_r)$ of the prior, which together with Equation 1 entails a factorization $p(\boldsymbol{\Gamma}|\mathcal{X}, \mathcal{S}^{(1:R)}) = \prod_{r=1}^{R} p(\boldsymbol{\gamma}_r|\mathcal{X}, \mathcal{S}^{(r)})$ of the posterior. Note that Equation 4 states that at test time we predict the reader r for which the marginal likelihood (that is, after integrating out the reader-specific model γ_r) of the test observations is highest. The next section discusses the reader-specific models $p(\mathbf{S}|\mathbf{X}, \boldsymbol{\gamma}_r)$ and prior distributions $p(\boldsymbol{\gamma}_r)$.

3 Probabilistic Model

The probabilistic model we employ follows the general structure proposed by Landwehr et al. (2014), but employs semiparametric density models and allows for fully Bayesian inference. To reduce notational clutter, let $\boldsymbol{\gamma} \in \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_R\}$ denote a particular reader-specific model, and let $\mathbf{X} \in \mathcal{X}$ denote a text. An eye movement pattern is a sequence $\mathbf{S} = ((s_1, d_1), \ldots, (s_T, d_T))$ of gaze fixations, consisting of a fixation position s_t (position in text that was fixated) and duration $d_t \in \mathbb{R}$ (length of fixation in milliseconds). In our experiments, individual sentences are presented in a single line on screen, thus we only model a horizontal gaze position $s_t \in \mathbb{R}$. We model $p(\mathbf{S}|\mathbf{X}, \boldsymbol{\gamma})$ as a dynamic process that successively generates fixation positions s_t and durations d_t in $\mathbf{S}$, reflecting how a reader generates a sequence of saccades in response to a text stimulus $\mathbf{X}$:

$$p(\mathbf{S}|\mathbf{X}, \boldsymbol{\gamma}) = p(s_1, d_1|\mathbf{X}, \boldsymbol{\gamma}) \prod_{t=2}^{T} p(s_t, d_t|s_{t-1}, \mathbf{X}, \boldsymbol{\gamma}),$$

where $p(s_t, d_t|s_{t-1}, \mathbf{X}, \boldsymbol{\gamma})$ models the generation of the next fixation position and duration given the old fixation position s_{t-1}. In the psychological literature, four different *saccade types* are distinguished: a reader can refixate the current word (*refixation*), fixate the next word in the text (*next word movement*), move the fixation to a word after the next word, that is, skip one or more words (*forward skip*), or regress to fixate a word occurring earlier in the text (*regression*), see, e.g., Heister et al. (2012). We observe empirically that for each saccade type, there is a characteristic distribution over saccade amplitudes and fixation durations, and that both approximately follow gamma distributions—see Figure 1. We therefore model $p(s_t, d_t|s_{t-1}, \mathbf{X}, \boldsymbol{\gamma})$ using a mixture over distributions for the four different saccade types. At each time t, the model first draws a saccade type $u_t \in \{1, 2, 3, 4\}$, and then draws a saccade amplitude a_t and fixation duration d_t from type-specific distributions $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\gamma})$ and $p(d|u_t, \boldsymbol{\gamma})$. More formally,

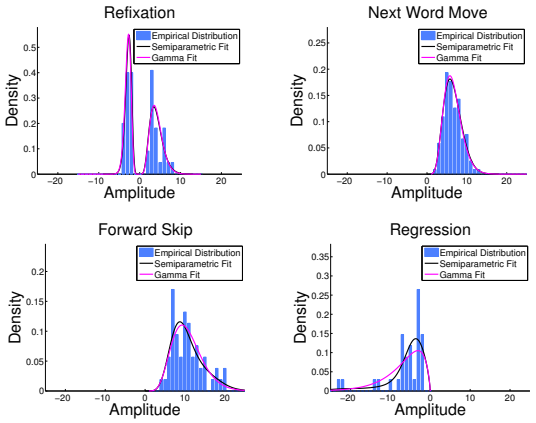

Figure 1: Empirical distributions of saccade amplitudes in training data for first individual, with fitted Gamma distributions and semiparametric distribution fits.

$$u_t \sim p(u|\boldsymbol{\pi}) \tag{7}$$
$$a_t \sim p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) \tag{8}$$
$$d_t \sim p(d|u_t, \boldsymbol{\delta}), \tag{9}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ is decomposed into components $\boldsymbol{\pi}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\delta}$. Afterwards, the model updates the fixation position according to $s_t = s_{t-1} + a_t$, concluding the definition of $p(s_t, d_t|s_{t-1}, \mathbf{X}, \boldsymbol{\gamma})$. Figure 2 shows a slice in the dynamical model.

The distribution $p(u|\boldsymbol{\pi})$ over saccade types (Equation 7) is multinomial with parameter vector $\boldsymbol{\pi} \in \mathbb{R}^4$. The distributions over amplitudes and durations (Equations 8 and 9) are modeled semiparametrically as discussed in the following subsections.

3.1 Model of Saccade Amplitudes

We first discuss the amplitude model $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha})$ (Equation 8). We first define a distribution $p(a|u_t, \boldsymbol{\alpha})$ over amplitudes for saccade type u_t, and subsequently discuss conditioning on the text $\mathbf{X}$ and old fixation position s_{t-1},

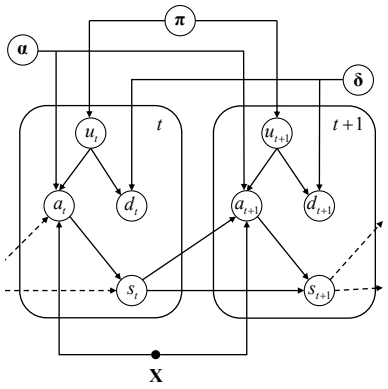

Figure 2: Plate notation of of a slice in the dynamic model.

leading to $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha})$. We define

$$p(a|u_t = 1, \boldsymbol{\alpha}) = \begin{cases} \mu\alpha_1(a) & : a > 0 \\ (1-\mu)\bar{\alpha}_1(-a) & : a \leq 0 \end{cases} \quad (10)$$

where μ is a mixture weight and α_1, $\bar{\alpha}_1$ are densities defining the distribution over positive and negative amplitudes for the saccade type *refixation*, and

$$p(a|u_t = 2, \boldsymbol{\alpha}) = \alpha_2(a) \quad (11)$$

$$p(a|u_t = 3, \boldsymbol{\alpha}) = \alpha_3(a) \quad (12)$$

$$p(a|u_t = 4, \boldsymbol{\alpha}) = \alpha_4(-a) \quad (13)$$

where $\alpha_2(a)$, $\alpha_3(a)$, and $\alpha_4(a)$ are densities defining the distribution over amplitudes for the remaining saccade types. Finally, the distribution

$$p(s_1|\mathbf{X}, \boldsymbol{\alpha}) = \alpha_0(s_1) \quad (14)$$

over the initial fixation position is given by another density function α_0. The variables $\mu, \alpha_0, \alpha_1, \bar{\alpha}_1$, α_2, α_3, and α_4 are aggregated into model component $\boldsymbol{\alpha}$. For resolving the most likely reader at test time (Equation 4), densities in $\boldsymbol{\alpha}$ will be integrated out under a prior based on Gaussian processes (Section 3.3) using MCMC inference (Section 4).

Given the old fixation position s_{t-1}, the text $\mathbf{X}$, and the chosen saccade type u_t, the amplitude is constrained to fall within a specific interval. For instance, for a refixation the amplitude has to be chosen such that the novel fixation position lies within the beginning and the end of the currently fixated word; a regression implies an amplitude that is negative and makes the novel fixation position lie before the beginning of the currently fixated word. These constraints imposed by the text structure define the conditional distribution $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha})$. More formally, $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha})$ is the distribution $p(a|u_t, \boldsymbol{\alpha})$ conditioned on $a \in [l, r]$, that is,

$$p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) = p(a|a \in [l, r], u_t, \boldsymbol{\alpha}),$$

where l and r are the minimum and maximum amplitude consistent with the constraints. Recall that for a distribution over a continuous variable x given by density $\alpha(x)$, the distribution over x conditioned on $x \in [l, r]$ is given by the truncated density

$$\alpha(x|x \in [l, r]) = \begin{cases} \frac{\alpha(x)}{\int_l^r \alpha(\bar{x})\mathrm{d}\bar{x}} & : x \in [l, r] \\ 0 & : x \notin [l, r]. \end{cases} \quad (15)$$

We derive $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha})$ by truncating the distributions given by Equations 10 to 13 to the minimum and maximum amplitude consistent with the current fixation position s_{t-1} and text $\mathbf{X}$. Let w_l° (w_r°) denote the position of the left-most (right-most) character of the currently fixated word, and let w_l^+, w_r^+ denote these positions for the next word in $\mathbf{X}$. Let furthermore $l^\circ = w_l^\circ - s_{t-1}$, $r^\circ = w_r^\circ - s_{t-1}$, $l^+ = w_l^+ - s_{t-1}$, and $r^+ = w_r^+ - s_{t-1}$. Then

$$p(a|u_t = 1, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) = \begin{cases} \mu\alpha_1(a|a \in [0, r^\circ]) : a > 0 \\ (1-\mu)\bar{\alpha}_1(-a|a \in [l^\circ, 0]) : a \leq 0 \end{cases} \quad (16)$$

$$p(a|u_t = 2, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) = \alpha_2(a|a \in [l^+, r^+]) \quad (17)$$

$$p(a|u_t = 3, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) = \alpha_3(a|a \in (r^+, \infty)) \quad (18)$$

$$p(a|u_t = 4, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) = \alpha_4(-a|a \in (-\infty, l^\circ)) \quad (19)$$

defines the appropriately truncated distributions.

3.2 Model of Fixation Durations

The model for fixation durations (Equation 9) is similarly specified by saccade type-specific densities,

$$p(d|u_t = u, \boldsymbol{\delta}) = \delta_u(d) \quad \text{for } u \in \{1, 2, 3, 4\} \quad (20)$$

and a density for the initial fixation durations

$$p(d_1|\mathbf{X}, \boldsymbol{\delta}) = \delta_0(d_1) \quad (21)$$

where $\delta_0, ..., \delta_4$ are aggregated into model component $\boldsymbol{\delta}$. Unlike saccade amplitude, the fixation duration is not constrained by the text structure and accordingly densities are not truncated. This concludes the definition of the model $p(\mathbf{S}|\mathbf{X}, \boldsymbol{\gamma})$.

3.3 Prior Distributions

The prior distribution over the entire model $\boldsymbol{\gamma}$ factorizes over the model components as

$$p(\boldsymbol{\gamma}|\lambda,\rho,\kappa) = \tag{22}$$
$$p(\boldsymbol{\pi}|\lambda)p(\mu|\rho)p(\bar{\alpha}_1|\kappa)\prod_{i=0}^{4}p(\alpha_i|\kappa)\prod_{i=0}^{4}p(\delta_i|\kappa)$$

where $p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\lambda)$ is a symmetric Dirichlet prior and $p(\mu) = \text{Beta}(\mu|\rho)$ is a Beta prior. The key challenge is to develop appropriate priors for the densities defining saccade amplitude ($p(\bar{\alpha}_1|\kappa), p(\alpha_i|\kappa)$) and fixation duration ($p(\delta_i|\kappa)$) distributions. Empirically, we observe that amplitude and duration distributions tend to be close to gamma distributions—see the example in Figure 1.

Our goal is to exploit the prior knowledge that distributions tend to be closely approximated by gamma distributions, but allow the model to deviate from the gamma assumption in case there is enough evidence in the data. To this end, we define a prior over densities that concentrates probability mass around the gamma family. For all densities $f \in \{\bar{\alpha}_1, \alpha_0, ..., \alpha_4, \delta_0, ..., \delta_4\}$, we employ identical prior distributions $p(f|\kappa)$. Intuitively, the prior is given by first drawing a density function from the gamma family and then drawing the final density from a Gaussian process (with covariance function κ) centered at this function. More formally, let

$$\mathcal{G}(x|\boldsymbol{\eta}) = \frac{\exp(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(x))}{\int \exp(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(x'))\mathrm{d}x'} \tag{23}$$

denote the gamma distribution in exponential family form, with sufficient statistics $\mathbf{u}(x) = (\log(x), x)^\mathsf{T}$ and parameters $\boldsymbol{\eta} = (\eta_1, \eta_2)$. Let $p(\boldsymbol{\eta})$ denote a prior over the gamma parameters, and define

$$p(f|\kappa) = \int p(\boldsymbol{\eta})p(f|\boldsymbol{\eta},\kappa)\mathrm{d}\boldsymbol{\eta} \tag{24}$$

where $p(f|\boldsymbol{\eta},\kappa)$ is given by drawing

$$g \sim \mathcal{GP}(0,\kappa) \tag{25}$$

from a Gaussian process prior $\mathcal{GP}(0,\kappa)$ with mean zero and covariance function κ, and letting

$$f(x) = \frac{\exp(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(x) + g(x))}{\int \exp(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(x') + g(x'))\mathrm{d}x'}. \tag{26}$$

Note that decreasing the variance of the Gaussian process means regularizing $g(x)$ towards zero, and therefore Equation 26 towards Equation 23. This concludes the specification of the prior $p(\boldsymbol{\gamma}|\lambda,\rho,\kappa)$.

The density model defined by Equations 24 to 26 draws on ideas from the large body of literature on GP-based density estimation, for example by Adams et al. (2009), Leonard (1978), or Tokdar et al. (2010), and semiparametric density estimation, e.g. as discussed by Yang (2009), Lenk (2003) or Hjort & Glad (1995). However, note that existing density estimation approaches are not applicable off-the-shelf as in our domain distributions are truncated differently at each observation due to constraints that arise from the way eye movements interact with the text structure (Equations 16 to 19).

4 Inference

To solve Equation 4, we need to integrate for each $r \in \mathcal{R}$ over the reader-specific model $\boldsymbol{\gamma}_r$. To reduce notational clutter, let $\boldsymbol{\gamma} \in \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_R\}$ denote a reader-specific model, and let $\mathcal{S} \in \{\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(R)}\}$ denote the eye movement observations of that reader on the training texts $\mathcal{X}$. We approximate

$$\int p(\bar{\mathcal{S}}|\bar{\mathcal{X}},\boldsymbol{\gamma})p(\boldsymbol{\gamma}|\mathcal{X},\mathcal{S})\mathrm{d}\boldsymbol{\gamma} \approx \frac{1}{K}\sum_{k=1}^{K}p(\bar{\mathcal{S}}|\bar{\mathcal{X}},\boldsymbol{\gamma}^{(k)})$$

by a sample $\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(K)}$ of models drawn by

$$\boldsymbol{\gamma}^{(k)} \sim p(\boldsymbol{\gamma}|\mathcal{X},\mathcal{S},\lambda,\rho,\kappa),$$

where $p(\boldsymbol{\gamma}|\mathcal{X},\mathcal{S},\lambda,\rho,\kappa)$ is the posterior as given by Equation 6 but with the dependence on the prior hyperparameters λ,ρ,κ made explicit. Note that with $\mathcal{X}$ and $\mathcal{S}$, all saccade types u_t are observed. Together with the factorizing prior (Equation 22), this means that the posterior factorizes according to

$$p(\boldsymbol{\gamma}|\mathcal{X},\mathcal{S},\lambda,\rho,\kappa) = p(\boldsymbol{\pi}|\mathcal{X},\mathcal{S},\lambda)p(\mu|\mathcal{X},\mathcal{S},\rho)$$
$$\cdot\, p(\bar{\alpha}_1|\mathcal{X},\mathcal{S},\kappa)\prod_{i=0}^{4}p(\alpha_i|\mathcal{X},\mathcal{S},\kappa)\prod_{i=0}^{4}p(\delta_i|\mathcal{X},\mathcal{S},\kappa)$$

as is easily seen from the graphical model in Figure 2. Obtaining samples $\boldsymbol{\pi}^{(k)} \sim p(\boldsymbol{\pi}|\mathcal{X},\mathcal{S})$ and $\mu^{(k)} \sim p(\mu|\mathcal{X},\mathcal{S})$ is straightforward because their prior distributions are conjugate to the likelihood terms. Let now $f \in \{\bar{\alpha}_1, \alpha_0, ..., \alpha_4, \delta_0, ..., \delta_4\}$

denote a particular density in the model. The posterior $p(f|\mathcal{X}, \mathcal{S}, \kappa)$ is proportional to the prior $p(f|\kappa)$ (Equation 24) multiplied by the likelihood of all observations that are generated by this density, that is, that are generated according to Equation 14, 16, 17, 18, 19, 20, or 21. Let $\mathbf{y} = (y_1, \ldots, y_{|\mathbf{y}|})^\mathsf{T} \in \mathbb{R}^{|\mathbf{y}|}$ denote the vector of all observations generated from density f, and let $\mathbf{l} = (l_1, \ldots, l_{|\mathbf{l}|})^\mathsf{T} \in \mathbb{R}^{|\mathbf{l}|}$, $\mathbf{r} = (r_1, \ldots, r_{|\mathbf{r}|})^\mathsf{T} \in \mathbb{R}^{|\mathbf{r}|}$ denote the corresponding left and right boundaries of the truncation intervals (again see Equations 14 to 21), where for densities that are not truncated we take $l_i = 0$ and $r_i = \infty$ throughout. Then the likelihood of the observations generated from f is

$$p(\mathbf{y}|f, \mathbf{l}, \mathbf{r}) = \prod_{i=1}^{|\mathbf{y}|} f(y_i|y_i \in [l_i, r_i]) \qquad (27)$$

and the posterior over f is given by

$$p(f|\mathcal{X}, \mathcal{S}, \kappa) \propto p(f|\kappa)p(\mathbf{y}|f, \mathbf{l}, \mathbf{r}). \qquad (28)$$

Note that $\mathbf{y}$, $\mathbf{l}$ and $\mathbf{r}$ are observable from $\mathcal{X}$, $\mathcal{S}$.

We obtain samples from the posterior given by Equation 28 from a Metropolis-Hastings sampler that explores the space of densities $f : \mathbb{R} \to \mathbb{R}$, generating density samples $f^{(1)}, ..., f^{(K)}$. A density f is given by a combination of gamma parameters $\boldsymbol{\eta} \in \mathbb{R}^2$ and function $g : \mathbb{R} \to \mathbb{R}$; specifically, f is obtained by multiplying the gamma distribution with parameters $\boldsymbol{\eta}$ by $\exp(g)$ and normalizing appropriately (Equation 26). During sampling, we explicitly represent a density sample $f^{(k)}$ by its gamma parameters $\boldsymbol{\eta}^{(k)}$ and function $g^{(k)}$. The proposal distribution of the Metropolis-Hastings sampler is

$$q(\boldsymbol{\eta}^{(k+1)}, g^{(k+1)}|\boldsymbol{\eta}^{(k)}, g^{(k)}) = p(g^{(k+1)}|\kappa)\mathcal{N}(\boldsymbol{\eta}^{(k+1)}|\boldsymbol{\eta}^{(k)}, \sigma^2\mathbf{I})$$

where $p(g^{(k+1)}|\kappa)$ is the probability of $g^{(k+1)}$ according to the GP prior $\mathcal{GP}(0, \kappa)$ (Equation 25), and $\mathcal{N}(\boldsymbol{\eta}^{(k+1)}|\boldsymbol{\eta}^{(k)}, \sigma^2\mathbf{I})$ is a symmetric proposal that randomly perturbs the old state $\boldsymbol{\eta}^{(k)}$ according to a Gaussian. In every iteration k a proposal $\boldsymbol{\eta}^\star, g^\star \sim q(\boldsymbol{\eta}, g|\boldsymbol{\eta}^{(k)}, g^{(k)})$ is drawn based on the old state $(\boldsymbol{\eta}^{(k)}, g^{(k)})$. The acceptance probability is $A(\boldsymbol{\eta}^\star, g^\star|\boldsymbol{\eta}^{(k)}, g^{(k)}) = \min(1, Q)$ with

$$Q = \frac{q(\boldsymbol{\eta}^{(k)}, g^{(k)}|\boldsymbol{\eta}^\star, g^\star)p(\boldsymbol{\eta}^\star)p(g^\star|\kappa)p(\mathbf{y}|f^\star, \mathbf{l}, \mathbf{r})}{q(\boldsymbol{\eta}^\star, g^\star|\boldsymbol{\eta}^{(k)}, g^{(k)})p(\boldsymbol{\eta}^{(k)})p(g^{(k)}|\kappa), p(\mathbf{y}|f^{(k)}, \mathbf{l}, \mathbf{r})}.$$

Here, $p(\boldsymbol{\eta}^\star)$ is the prior probability of gamma parameters $\boldsymbol{\eta}^\star$ (Section 3.3) and $p(\mathbf{y}|f^\star, \mathbf{l}, \mathbf{r})$ is given by Equation 27 where $f^\star$ is obtained from $\boldsymbol{\eta}^\star$, $g^\star$ according to Equation 26.

To compute the likelihood terms $p(\mathbf{y}|f^{(k)}, \mathbf{l}, \mathbf{r})$ (Equation 27) and also to compute the likelihood of test data under a model (Equation 5), the density $f : \mathbb{R} \to \mathbb{R}$ needs to be evaluated. According to Equation 26, f is represented by parameter vector $\boldsymbol{\eta}$ together with the nonparametric function $g : \mathbb{R} \to \mathbb{R}$. As usual when working with distributions over functions in a Gaussian process framework, the function g only needs to be represented at those points for which we need to evaluate it. Clearly, this includes all observations of saccade amplitudes and fixation durations observed in the training and test set. However, we also need to evaluate the normalizer in Equation 26, and (for $f \in \{\alpha_1, \bar{\alpha}_1, \alpha_2, \alpha_3, \alpha_4\}$) the additional normalizer required when truncating the distribution (see Equation 15). As these integrals are one-dimensional, they can be solved relatively accurately using numerical integration; we use 2-point Newton-Cotes quadrature. Newton-Cotes integration requires the evaluation (and thus representation) of g at an additional set of equally spaced supporting points.

When the set of test observations $\bar{\mathcal{S}}, \bar{\mathcal{X}}$ is large, the need to evaluate $p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \boldsymbol{\gamma}^{(k)})$ for all $\boldsymbol{\gamma}_k$ and all test observations leads to computational challenges. In our experiments, we use a heuristic to reduce computational load. While generating samples, densities are only represented at the training observations and the supporting points needed for Newton-Cotes integration. We then estimate the mean of the posterior by $\hat{\boldsymbol{\gamma}} = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\gamma}^{(k)}$, and approximate $\frac{1}{K}\sum_{k=1}^K p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \boldsymbol{\gamma}^{(k)}) \approx p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \hat{\boldsymbol{\gamma}})$. To evaluate $p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \hat{\boldsymbol{\gamma}})$, we infer the approximate value of the density $\hat{\boldsymbol{\gamma}}$ at a test observation by linearly interpolating based on the available density values at the training observations and supporting points.

5 Empirical Study

We conduct a large-scale study of biometric identification performance using the same setup as discussed by Landwehr et al. (2014) but a much larger set of individuals (251 rather than 20).

Eye movement records for 251 individuals are

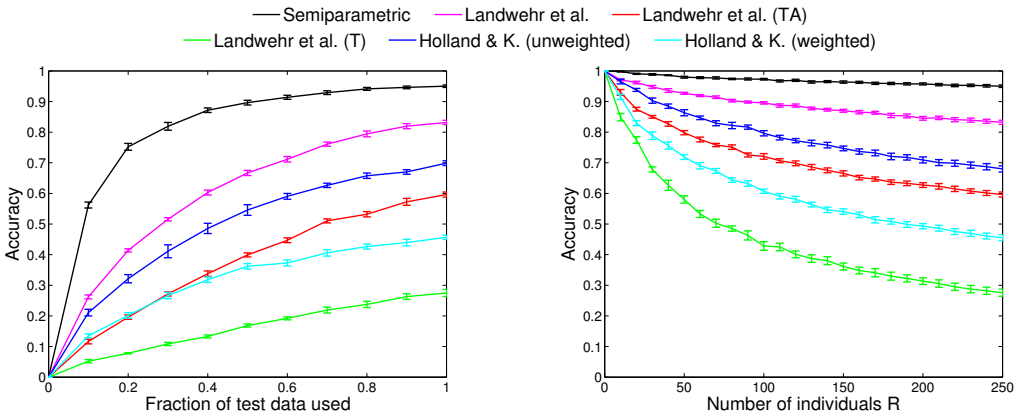

Figure 3: Multiclass accuracy over number of test observations (left) and number of individuals R (right) with standard errors.

Method	Accuracy
Semiparametric	0.9502 ± 0.0130
Semiparametric (TD)	0.8853 ± 0.0142
Semiparametric (TA)	0.7717 ± 0.0361
Landwehr et al.	0.8319 ± 0.0218
Landwehr et al. (TA)	0.5964 ± 0.0262
Landwehr et al. (T)	0.2749 ± 0.0369
Holland & K. (unweighted)	0.6988 ± 0.0241
Holland & K. (weighted)	0.4566 ± 0.0220

Table 1: Multiclass identification accuracy $\pm$ standard error.

obtained from an EyeLink II system with a 500-Hz sampling rate (SR Research, Ongoode, Ontario, Canada) while reading sentences from the *Potsdam Sentence Corpus* (Kliegl et al., 2006). There are 144 sentences in the corpus, which we split into equally sized sets of training and test sentences. Individuals read between 100 and 144 sentences, the training (testing) observations for one individual are the observations on those sentences in the training (testing) set of sentences that the individual has read. Results are averaged over 10 random train-test splits. Each sentence is shown as a single line on the screen.

We study the semiparametric model discussed in Section 3 with MCMC inference as presented in Section 4 (denoted *Semiparametric*[1]). We employ a squared exponential covariance function $\kappa(x, x') = \alpha \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$, where the multiplicative constant α is tuned on the training data by cross-validation and the bandwidth σ is set to the average distance between points in the training data. The Beta and Dirichlet parameters λ and ρ are set to one (Laplace smoothing), the prior $p(\boldsymbol{\eta})$ for the Gamma parameters is uninformative. We use backoff-smoothing as discussed by Landwehr et al. (2014). We initialize the sampler with the maximum-likelihood Gamma fit and perform 10000 sampling iterations, 5000 of which are burn-in iterations. As a baseline, we study the model by Landwehr et al. (2014) (*Landwehr et al.*) and simplified versions proposed by them that only use saccade type and amplitude (*Landwehr et al. (TA)*) or saccade type (*Landwehr et al. (T)*). We also study the weighted and unweighted version of the feature-based model of Holland & Komogortsev (2012) with a feature set adapted to the Potsdam Sentence Corpus data as described in Landwehr et al. (2014).

We note that there are two recent extensions of the feature-based model (by Rigas et al. (2016) and Abdulin & Komogortsev (2015)) that are unfortunately not applicable in our empirical setting but might yield improved results in other scenarios. Rigas et al. (2016) study a model that is focused on representing reader-specific differences in saccadic vigor and acceleration, which are both derived from the dynamics of saccadic velocity. In the preprocessed data set that we use, saccadic velocities are not available, therefore we do not make use of velocities in our model and cannot easily compare against their model. Abdulin & Komogortsev (2015) study a model that is based on features that relate eye move-

[1] An implementation is available at `github.com/abdelwahab/SemiparametricIdentification`

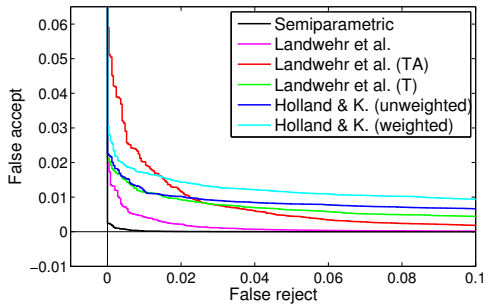

Figure 4: False-accept over false-reject rate when varying τ.

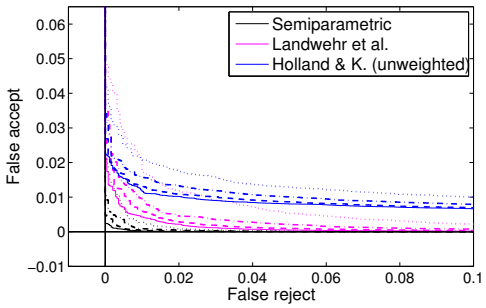

Figure 5: False-accept over false-reject rate when using 40% (dotted), 60% (dashed-dotted), 80% (dashed), and 100% (solid) of test observations, for selected subset of methods.

ments to the 2D text structure, that is, to the way words are arranged into lines in a text. As in our empirical study each sentence is presented as a single line on screen, this 2D structure does not exist. Moreover, Abdulin & Komogortsev (2015) only report accuracy improvements for their method in a setting where individuals have to be identified in the future based on data collected in the past (*aging test*), which is not the focus of our study.

We first study multiclass identification accuracy. All test observations of one particular individual constitute one test example; the task is to infer the individual that has generated these test observations. Multiclass identification accuracy is the fraction of cases in which the correct individual is identified. Table 1 shows multiclass identification accuracy for all methods, including variants of *Semiparametric* discussed below. We observe that *Semiparametric* outperforms *Landwehr et al.*, reducing the error by more than a factor of three. Consistent with results reported in Landwehr et al. (2014), *Holland & K. (unweighted)* is less accurate than *Landwehr et al.*, but more accurate than the simplified variants. We next study how the amount of data available at test time—that is, the amount of time we can observe a reader before having to make a decision—influences accuracy. Figure 3 (left) shows identification accuracy as a function of the fraction of test data available, obtained by randomly removing a fraction of sentences from the test set. We observe that identification accuracy steadily improves with more test observations for all methods. Figure 3 (right) shows identification accuracy when varying the number R of individuals that need to be distinguished. We randomly draw a subset of R individuals from the set

Method	Area under curve
Semiparametric	0.0000119
Semiparametric (TD)	0.0000821
Semiparametric (TA)	0.0001833
Landwehr et al.	0.0001743
Landwehr et al. (TA)	0.0010371
Landwehr et al. (T)	0.0017040
Holland & K. (unweighted)	0.0027853
Holland & K. (weighted)	0.0039978

Table 2: Area under the curve in binary classification setting.

of 251 individuals, and perform identification based on only these individuals. Results are averaged over 10 such random draws. As expected, accuracy improves if fewer individuals need to be distinguished.

We next study a binary setting in which for each individual and each set of test observations a decision has to be made whether or not the test observations have been generated by that individual. This setting more closely matches typical use cases for the deployment of a biometric system. Let $\bar{\mathcal{X}}$ denote the text being read at test time, and let $\bar{\mathcal{S}}$ denote the observed eye movement sequences. Our model infers for each reader $r \in \mathcal{R}$ the marginal likelihood $p(\bar{\mathcal{S}}|r, \bar{\mathcal{X}}, \mathcal{X}, \mathcal{S}^{(1:R)})$ of the eye movement observations under the reader-specific model (Equation 3). The binary decision is made by dividing this marginal likelihood by the average marginal likelihood assigned to the observations by all reader-specific models, and comparing the result to a threshold τ. Figure 4 shows the fraction of false accepts as a function of false rejects as the threshold τ is varied, averaged over all individuals. The *Landwehr et al.* model and variants also assign a

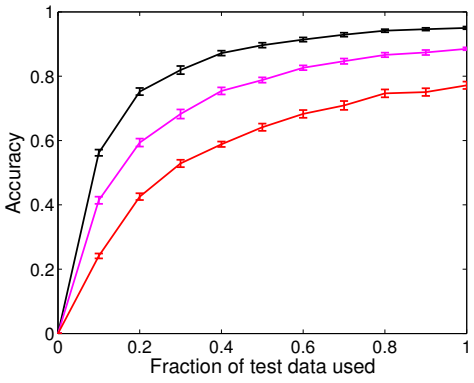

Figure 6: Multiclass accuracy over number of test observations with standard errors for *Semiparametric* variants.

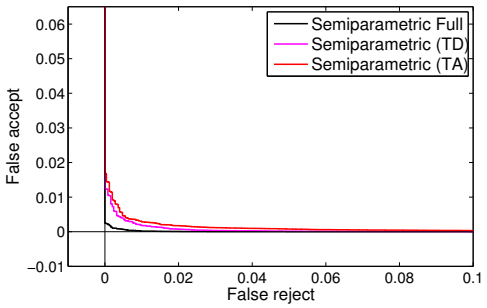

Figure 7: False-accept over false-reject rate when varying τ for the *Semiparametric* variants.

reader-specific likelihood to novel test observations; we compute the same statistics again by normalizing the likelihood and comparing to a threshold τ. Finally, *Holland & K. (unweighted)* and *Holland & K. (weighted)* compute a similarity measure for each combination of individual and set of test observations, which we normalize and threshold analogously. We observe that *Semiparametric* accomplishes a false-reject rate of below 1% at virtually no false accepts; *Landwehr et al.* and variants tend to perform better than *Holland & K. (unweighted)* and *Holland & K. (weighted)*. Table 2 shows the error under the curve for the experiment shown in Figure 4, as well as for variants of *Semiparametric* discussed below.

We finally study the contribution of the individual model components for saccade type, saccade amplitude, and fixation duration (see Figure 2) by removing the corresponding model components, as in Landwehr et al. (2014). By *Semiparametric (TD)* we denote a variant of *Semiparametric* in which the variable a_t and the corresponding distribution is removed, that is, only the distribution over the saccade type and duration is modeled. *Semiparametric (TA)* denotes a variant in which the variable d_t and the corresponding distribution is removed. Figure 6 shows identification accuracy as a function of the fraction of test data available for model variants *Semiparametric (TD)* and *Semiparametric (TA)* in comparison to *Semiparametric*; results for these variants are also included in Table 1. Figure 7 shows the fraction of false accepts as a function of false rejects in the binary classification setting discussed above for these two model variants; Table 2 includes area under the curve results for the experiment shown in Figure 7. We observe that accuracy is substantially reduced when removing any model component. Note that if both the amplitude and duration components of the model are removed, it becomes identical to the model *Landwehr et al. (T)*.

Training the joint model for all 251 individuals takes 46 hours on a single eight-core CPU (Intel Xeon E5520, 2.27GHz); predicting the most likely individual to have generated a set of 72 test sentences takes less than 2 seconds.

6 Conclusions

We have studied the problem of identifying readers unobtrusively during reading of arbitrary text. For fitting reader-specific distributions, we employ a Bayesian semiparametric approach that infers densities under a Gaussian process prior centered at the gamma family of distributions, striking a balance between robustness to sparse data and modeling flexibility. In an empirical study with 251 individuals, the model was shown to reduce identification error by more than a factor of three compared to earlier approaches to reader identification proposed by Landwehr et al. (2014) and Holland & Komogortsev (2012).

Acknowledgements

We gratefully acknowledge support from the German Research Foundation (DFG), grant LA 3270/1-1.

References

Evgeniy Abdulin and Oleg Komogortsev. 2015. Person verification via eye movement-driven text reading model. In *Proceedings of the Sixth International Conference on Biometrics: Theory, Applications and Systems*.

Ryan P. Adams, Iain Murray, and David J.C. MaxKay. 2009. Gaussian process density sampler. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems*.

Roman Bednarik, Tomi Kinnunen, Andrei Mihaila, and Pasi Fränti. 2005. Eye-movements as a biometric. In *Proceedings of the 14th Scandinavian Conference on Image Analysis*.

W. Robert Dixon. 1951. Studies in the psychology of reading. In W. S. Morse, P. A. Ballantine, and W. R. Dixon, editors, *Univ. of Michigan Monographs in Education No. 4*. Univ. of Michigan Press.

Ralf Engbert, Antje Nuthmann, Eike M. Richter, and Reinhold Kliegl. 2005. SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777–813.

Tadayoshi Hara, Daichi Mochihashi, Yoshino Kano, and Akiko Aizawa. 2012. Predicting word fixations in text with a CRF model for capturing general reading strategies among readers. In *Proceedings of the First Workshop on Eye-Tracking and Natural Language Processing*.

Julian Heister, Kay-Michael Würzner, and Reinhold Kliegl. 2012. Analysing large datasets of eye movements during reading. In James S. Adelman, editor, *Visual word recognition. Vol. 2: Meaning and context, individuals and development*, pages 102–130.

Nils L. Hjort and Ingrid K. Glad. 1995. Nonparametric density estimation with a parametric start. *The Annals of Statistics*, 23(3):882–904.

Corey Holland and Oleg V. Komogortsev. 2012. Biometric identification via eye movement scanpaths in reading. In *Proceedings of the 2011 International Joint Conference on Biometrics*.

Edmund B. Huey. 1908. *The psychology and pedagogy of reading*. Cambridge, Mass.: MIT Press.

Pawel Kasprowski and Jozef Ober. 2004. Eye movements in biometrics. In *Proceedings of the 2004 International Biometric Authentication Workshop*.

Reinhold Kliegl, Antje Nuthmann, and Ralf Engbert. 2006. Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, 135(1):12–35.

Oleg V. Komogortsev, Sampath Jayarathna, Cecilia R. Aragon, and Mechehoul Mahmoud. 2010. Biometric identification via an oculomotor plant mathematical model. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*.

Niels Landwehr, Sebastian Arzt, Tobias Scheffer, and Reinhold Kliegl. 2014. A model of individual differences in gaze control during reading. In *Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing*.

Peter J. Lenk. 2003. Bayesian semiparametric density estimation and model verification using a logistic-Gaussian process. *Journal of Computational and Graphical Statistics*, 12(3):548–565.

Tom Leonard. 1978. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society*, 40(2):113–146.

Franz Matties and Anders Søgaard. 2013. With blinkers on: robust prediction of eye movements across readers. In *Proceedings of the 2013 Conference on Empirical Natural Language Processing*.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Erik D. Reichle, Alexander Pollatsek, Donald L. Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological Review*, 105(1):125–157.

Ioannis Rigas, George Economou, and Spiros Fotopoulos. 2012a. Biometric identification based on the eye movements and graph matching techniques. *Pattern Recognition Letters*, 33(6).

Ioannis Rigas, George Economou, and Spiros Fotopoulos. 2012b. Human eye movements as a trait for biometrical identification. In *Proceedings of the IEEE 5th International Conference on Biometrics: Theory, Applications and Systems*.

Ioannis Rigas, Oleg Komogortsev, and Reza Shadmehr. 2016. Biometric recognition via eye movements: Saccadic vigor and acceleration cues. *ACM Transaction on Applied Perception*, 13(2):1–21.

Surya T. Tokdar, Yu M. Zhuy, and Jayanta K. Ghoshz. 2010. Bayesian density regression with logistic gaussian process and subspace projection. *Bayesian Analysis*, 5(2):319–344.

Ying Yang. 2009. Penalized semiparametric density estimation. *Statistics and Computing*, 19(1):355–366.

Youming Zhang and Martti Juhola. 2012. On biometric verification of a user by means of eye movement data mining. In *Proceedings of the 2nd International Conference on Advances in Information Mining and Management*.

Quantile Layers: Statistical Aggregation in Deep Neural Networks for Eye Movement Biometrics

Ahmed Abdelwahab ✉ and Niels Landwehr

Leibniz Institute of Agricultural Engineering and Bioeconomy e.V. (ATB), Potsdam, Germany {AAbdelwahab,NLandwehr}@atb-potsdam.de

Abstract. Human eye gaze patterns are highly individually characteristic. Gaze patterns observed during the routine access of a user to a device or document can therefore be used to identify subjects unobtrusively, that is, without the need to perform an explicit verification such as entering a password. Existing approaches to biometric identification from gaze patterns segment raw gaze data into short, local patterns called saccades and fixations. Subjects are then identified by characterizing the distribution of these patterns or deriving hand-crafted features for them. In this paper, we follow a different approach by training deep neural networks directly on the raw gaze data. As the distribution of short, local patterns has been shown to be particularly informative for distinguishing subjects, we introduce a parameterized and end-to-end learnable statistical aggregation layer called the *quantile layer* that enables the network to explicitly fit the distribution of filter activations in preceeding layers. We empirically show that deep neural networks with quantile layers outperform existing probabilistic and feature-based methods for identifying subjects based on eye movements by a large margin.

Keywords: eye movements · deep learning · biometry.

1 Introduction

Human visual perception is a fundamentally active process. We are not simply exposed to an incoming flow of visual sensory data, but rather actively control the visual input by continuously performing eye movements that direct the gaze focus to those points in space that are estimated to be most informative. The interplay between visual information processing and gaze control has been extensively studied in cognitive psychology, as it constitutes an important example of the link between cognitive processing and motor control [9, 19].

One insight from existing studies in psychology is that the resulting gaze patterns are highly individually characteristic [22, 23]. It is therefore possible to identify subjects based on their observed gaze patterns with high accuracy, and the use of gaze patterns as a biometric feature has been widely studied. Approaches for using gaze patterns for identification can be divided into two groups. One group of methods uses an active challenge-response protocol, that

is, identification is based on eye movements in response to an artificial visual stimulus [13, 25]. This has the disadvantage that additional time and effort of a user is required in order to confirm her identity. In the second group of methods, biometric identification is based on gaze patterns observed during the routine access of a user to a device or document [17, 26]. This way the identity can be confirmed unobtrusively, without requiring reaction to a specific challenge protocol. If the observed gaze patterns are unlikely to be generated by an authorized individual, access can be terminated or an additional verification requested.

Existing approaches for identifying subjects from gaze patterns mostly segment the raw eye gaze data into fixations (short periods of time in which the gaze is relatively stable) and saccades (rapid movements of the gaze to a new fixation position). They then either use probabilistic models that characterize the distribution of saccades and fixations [17, 1, 20], or hand-crafted statistical features that characterize different properties of saccades such as lengths, velocities, or accelerations [12, 26, 7]. In this paper, we follow a different approach by training deep neural networks on the raw gaze position data, without segmenting gaze movements into saccades and fixations or applying handcrafted aggregate features. However, we take inspiration from existing probabilistic approaches, which have shown that the distribution of local, short-term patterns in gaze movements such as saccades and fixations can be highly characteristic for different individuals. We therefore design neural network architectures that can extract such local patterns and characterize their distribution.

More specifically, we introduce a parameterized and end-to-end learnable statistical aggregation layer called the *quantile layer* that enables the network to explicitly fit the distribution of filter activations in preceeding layers. We design network architectures in which stacked 1D-convolution layers extract local, short-term patterns from eye movement sequences. The quantile layer characterizes the distribution of these patterns by approximating the *quantile function*, that is, the inverse cumulative distribution function, of the activations of the filters across the time series of gaze movements. The quantile function is approximated by sampling the empirical quantile function of the activations at a set of points, which are trainable model parameters. Natural special cases of the quantile layer are global maximum pooling and global median pooling; median pooling will approximate average pooling if filter activations are approximately symmetric. The proposed quantile layer can thus be seen as an extension of standard global pooling layers that retains more information about the distribution of activations than the average or maximum. In the same way as standard global pooling layers, the quantile layer aggregates over the entire sequence, enabling the model to work with variable-length sequences. By learning the sampling points, the model can focus on those parts of the distribution function that are most discriminative for identification. Using a piecewise linear approximation to the empirical quantile function makes the layer fully differentiable; models can thus be trained end-to-end using gradient descent. We empirically show that deep neural networks using quantile layers outperform existing probabilistic and feature-based approaches for identification based on gaze movements by a large margin.

Unobtrusive biometric identification has been most extensively studied based on gaze patterns during reading. In this paper, we study biometric eye gaze models for arbitrary non-text input. We specifically use data from the *dynamic images and eye movements* (DIEM) project, a large-scale data collection effort during which gaze movements of over 200 participants each watching a subset of 84 video sequences were recorded [21]. This data is approximately representative of scenarios where a user is not reading text (e.g., watching a live stream from a security camera), broadening the application range of gaze-based biometrics.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 introduces the quantile layer, Section 4 discusses deep neural network architectures for eye gaze biometrics. We empirically study identification accuracy of the proposed methods and different baselines in Section 5.

2 Related Work

Biometric identification from eye gaze patterns observed as a response to a specific stimulus has been studied extensively. The stimulus can for example be a moving [13, 16, 18, 31] or fixed [2] dot on a monitor, or a specific image stimulus [25]. More recently, unobtrusive biometric identification based on gaze patterns observed during the routine access of a user to a device or document has been studied. This approach has the advantage that no additional time and attention of a user are needed for identification, because gaze patterns are generated on material that is viewed anyway. Most unobtrusive approaches are based on observing eye movements of subjects generated while reading text [11, 1, 26], but identification based on eye movements generated while viewing non-text input has also been studied [15].

Existing approaches for biometric identification (with the exception of the work by Kinnunen et al. [15], see below) first segment the observed eye movement data into fixations (periods of little gaze movement during which the visual content at the current position is processed) and saccades (short, ballistic movements that relocate the gaze to a new fixation position). One approach that has been widely studied in the literature is to derive hand-crafted features of these saccades and fixations that are believed to be characteristic for individual subjects. Holland and Komogortsev have studied relatively simple features such as average fixation duration, average saccade amplitude and average saccade velocity [11, 12]. This line of work was later extended to more complex features such as saccadic vigor, acceleration, or the so-called *main sequence* feature [26, 7]. Subjects are then identified by matching the features of observed eye gaze sequences generated by an unknown individual to those of known individuals, using for example shortest distance[11], statistical tests [12, 26], or an RBF classifier [7].

Another popular approach is to use probabilistic models that characterize user-specific distributions over saccades and fixations. Landwehr et al. [17] have studied simple parametric models based on the Gamma family. Abdelwahab et al. [1] have studied semiparametric models in which the identity of a user is inferred by Bayesian inference based on Metropolis-Hastings sampling under

a Gaussian process prior. Makowski et al. [20] study a discriminative model that takes into account lexical features of fixated words, such as word frequency and word lengths, and show that this can further increase identification accuracy from gaze patterns obtained during reading. The approach discussed by Kinnunen et al. [15] also uses a probabilistic approach, by fitting a Gaussian mixture model to the distribution of angles between successive gaze positions. Unlike the approaches discussed above, Kinnunen et al. do not segment the eye signal into fixations and saccades, but rather use all recorded gaze positions. Our work differs from these existing approaches to biometric identification from gaze patterns in that we train deep neural networks on the raw eye gaze to distinguish between different subjects. We show empirically that this leads to large gains in identification accuracy compared to existing feature-based and probabilistic approaches, including the model by Kinnunen et al. [15].

The quantile layer we propose as a more expressive statistical aggregation layer than standard global pooling is related to the learnable *histogram layers* proposed by Wang et al. [30] and Sedighi and Fridrich [27]. Histogram layers are also fully differentiable, parameterized statistical aggregation layers. They characterize the distribution of values in the input to the layer in terms of an approximation to a histogram, in which bin centers and bin widths are learnable parameters. Wang et al. [30] use linear approximations to smoothen the sharp edges in a traditional histogram function and enable gradient flow. Sedighi and Fridrich [27] use Gaussian kernels as a soft, differentiable approximation to histogram bins. The histogram layers proposed by Wang et al. [30] and Sedighi and Fridrich [27] directly approximate the probability density of the input values, while the quantile layer we propose approximates the cumulative distribution function. The quantile layer also naturally generalizes maximum pooling and median pooling, while the histogram layers do not directly relate to standard pooling operations. We use architectures based on the histogram layers of Wang et al. [30] and Sedighi and Fridrich [27] as baselines in our empirical study.

Finally, Couture et al. [5] have recently studied quantiles as a method to aggregate instance-level predictions when training deep multi-instance neural networks for detecting tumor type from tissue images. In their application, images are represented as bags of subimages, and predictions on individual subimages are combined into a bag prediction based on the quantile function.

3 The Quantile Layer

This section introduces the quantile layer, a parameterized and end-to-end learnable layer for characterizing the distribution of filter activations in a preceeding convolution layer. This layer will be a central component in the deep neural network architectures for eye gaze biometrics that we develop in the next section.

The gaze movement data we study is a discrete time series of 2D-coordinates that indicate the current focus point of the gaze on a plane (e.g., a monitor). The discrete time series is obtained by sampling the continuous gaze movements at a regular frequency, and can be observed using standard eye tracking devices.

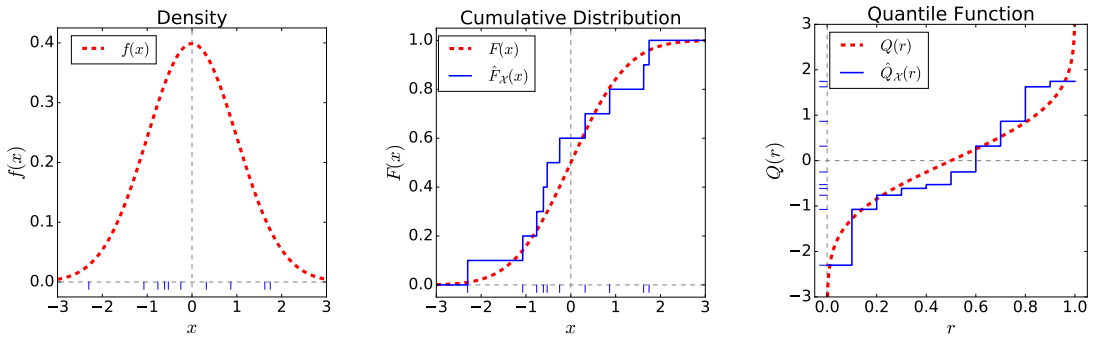

Fig. 1. Density function, cumulative distribution function, and quantile function (dashed lines) with empirical counterparts (solid lines) for a normally distributed variable $x \sim \mathcal{N}(0, 1)$. Tick marks at zero line show a sample from the distribution.

Existing approaches for user identification from eye movements first preprocess the raw signal into two kinds of short, local patterns: saccades (rapid movements, characterized by their amplitude) and fixations (periods of almost constant gaze position, characterized by their duration). They then distinguish users based on their distribution of saccade amplitudes and fixation durations (and possibly other local features). This is done either by computing aggregate features [11, 12, 26] or by fitting parametric or semiparametric probabilistic models to the observed distributions [17, 1, 20]. The key insight from this existing work is that the most informative feature for identification is the distribution of short, local gaze patterns seen in a particular sequence. In contrast, long-term dependencies in the time series will be less informative, as these are more likely to be a function of the visual input than the identity of the viewer.

Motivated by these observations in earlier work, we study network architectures that consists of a deep arrangement of 1D-convolution filters, which extract local, short-term patterns from the raw gaze signal, followed by the quantile layer whose output characterizes the distribution of these patterns. We design the quantile layer in such a way that it naturally generalizes global maximum, median, and minimum pooling. As we assume that the distribution of short-term patterns is most informative, we use standard non-dilated convolution operations, rather than dilated convolution operations which have recently been used for modeling more long-term patterns in time series, for example for audio data [29].

Let x denote a real-valued random variable whose distribution is given by the probability density function $f(x)$. The distribution of x can be expressed in different forms: by the density function $f(x)$, by the cumulative distribution function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = \int_{-\infty}^{x} f(z)\mathrm{d}z, \tag{1}$$

or by the *quantile function* $Q : (0, 1) \to \mathbb{R}$ defined by

$$Q(r) = \inf\{x \in \mathbb{R} : r \leq F(x)\} \tag{2}$$

where inf denotes the infimum and $(0, 1) \subset \mathbb{R}$ the open interval from zero to one. The quantile function Q is characterized by $p(x \leq Q(r)) = r$. That is, the quantile function yields the value $Q(r) \in \mathbb{R}$ such that all values of the random variable x smaller than $Q(r)$ together account for probability mass r. If the cumulative distribution function F is continuous and strictly monotonically increasing, which it will be if the density function $f(x)$ is continuous and positive everywhere on $\mathbb{R}$, the quantile function Q is simply the inverse of the cumulative distribution function, $Q = F^{-1}$. Figure 1 visualizes the relationship between density, cumulative distribution, and quantile functions for a standard normally distributed variable $x \sim \mathcal{N}(0, 1)$.

If $\mathcal{X} = \{x_1, ..., x_n\}$ with $x_i \sim p(x)$ denotes a sample of the random variable x, the empirical cumulative distribution function $\hat{F}_{\mathcal{X}} : \mathbb{R} \to [0, 1]$ is a non-parametric estimator of the cumulative distribution function F. It is given by

$$\hat{F}_{\mathcal{X}}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x) \tag{3}$$

where

$$I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x. \end{cases} \tag{4}$$

In analogy to the empirical distribution function, the empirical quantile function $\hat{Q}_{\mathcal{X}} : (0, 1] \to \mathbb{R}$ is a non-parametric estimator of the quantile function Q. It is defined by

$$\hat{Q}_{\mathcal{X}}(r) = \inf\{x \in \mathbb{R} : r \leq \hat{F}_{\mathcal{X}}(x)\}. \tag{5}$$

Figure 1 visualizes the empirical cumulative distribution function $\hat{F}(x)$ and the empirical quantile function $\hat{Q}(r)$ together with a set of samples for a standard normally distributed variable. For sufficiently large sample size n, the empirical quantile function faithfully characterizes the distribution of x in the following sense. According to the Glivenko-Cantelli theorem, $\hat{F}_{\mathcal{X}}$ uniformly converges to the true cumulative distribution function F,

$$\sup_{x \in \mathbb{R}} |\hat{F}_{\mathcal{X}}(x) - F(x)| \xrightarrow{a.s.} 0 \tag{6}$$

[28], where we use $\xrightarrow{a.s.}$ to denote almost sure convergence in the sample size n. For all $r \in (0, 1)$ this implies almost sure convergence of $\hat{Q}_{\mathcal{X}}(r)$ to $Q(r)$,

$$|\hat{Q}_{\mathcal{X}}(r) - Q(r)| \xrightarrow{a.s.} 0 \tag{7}$$

provided that Q is continuous at r [24]. The empirical quantile function thus faithfully estimates the quantile function in the limit. Finally, the quantile function Q determines the distribution over x, that is, for a given quantile function Q there is a unique cumulative distribution function F such that Equation 2 is satisfied [6].

Let $\pi : \{1, ..., n\} \to \{1, ..., n\}$ denote a permutation that sorts the sample in ascending order, that is, $x_{\pi(i)} \leq x_{\pi(i+1)}$ for $i \in \{1, ..., n - 1\}$. Then

$$\hat{Q}_{\mathcal{X}}(r) = x_{\pi(k)} \tag{8}$$

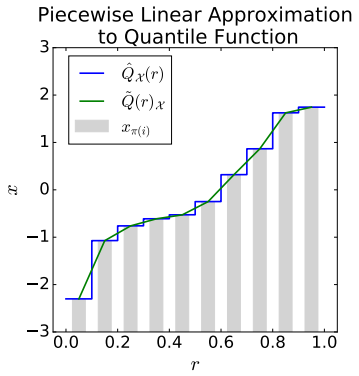

Fig. 2. Empirical quantile function, sorted samples, and piecewise linear approximation to the empirical quantile function. The set of samples is identical to that in Figure 1.

for the unique $k \in \mathbb{N}$ fulfilling the condition

$$\frac{k-1}{n} < r \leq \frac{k}{n}. \tag{9}$$

That is, the empirical quantile function $\hat{Q}_{\mathcal{X}}(r)$ can be computed by sorting the samples in ascending order, and returning the sample at position $\lceil r \cdot n \rceil$, where for $x \in \mathbb{R}$ we use $\lceil x \rceil$ to denote the smallest integer larger than or equal to x. This is visualized in Figure 2, where the ordered samples $x_{\pi(1)}, ..., x_{\pi(n)}$ are shown as a bar plot together with $\hat{Q}_{\mathcal{X}}$.

We will also work with a piecewise linear approximation $\tilde{Q}_{\mathcal{X}}$ to the empirical quantile function $\hat{Q}_{\mathcal{X}}$, as shown in Figure 2. This function is defined on the interval $[\frac{1}{2n}, 1 - \frac{1}{2n}]$ by $\tilde{Q}_{\mathcal{X}}(\frac{2k-1}{2n}) = \hat{Q}_{\mathcal{X}}(\frac{2k-1}{2n})$ for $k \in \{1, ..., n\}$ and by being piecewise linear in between. The piecewise linear approximation is needed in order to make the quantile layer that we introduce below fully differentiable. Note that $\tilde{Q}_{\mathcal{X}}$ will return the minimum, median, and maximum of the set of samples as special cases. Equation 8 implies $\tilde{Q}_{\mathcal{X}}(\frac{1}{2n}) = \min\{x_1, ..., x_n\}$, $\tilde{Q}_{\mathcal{X}}(0.5) = \text{med}\{x_1, ..., x_n\}$, and $\tilde{Q}_{\mathcal{X}}(1 - \frac{1}{2n}) = \max\{x_1, ..., x_n\}$.

We now define the *quantile layer* as the operation of sampling the piecewise linear approximation $\tilde{Q}_{\mathcal{X}}$ to the empirical quantile function $\hat{Q}_{\mathcal{X}}$ for a set $\mathcal{X}$ of incoming filter activations. The quantile layer takes as input the output of a convolution layer, and outputs a set of features in which the temporal dimension has been aggregated out. The input to the quantile layer is thus a matrix $\mathbf{Z} \in \mathbb{R}^{T \times K}$ of activations, where K is the number of filters and T the temporal dimension in the preceding convolution layer. The output of the quantile layer is a matrix $\mathbf{Y} \in \mathbb{R}^{K \times M}$, where M is a hyperparameter that determines at how many points $\hat{Q}_{\mathcal{X}}$ is sampled. Let $z_{t,k}$ denote the element at row t and column k of $\mathbf{Z}$, and $y_{k,m}$ denote the element at row k and column m of $\mathbf{Y}$. Then the outputs $y_{k,m}$ of the layer are defined by

$$y_{k,m} = \tilde{Q}_{\mathcal{X}_k}\left(\sigma(\alpha_{k,m})\frac{T-1}{T} + \frac{1}{2T}\right) \tag{10}$$

where $\mathcal{X}_k = \{z_{t,k} | 1 \leq t \leq T\}$ is the set of activations of filter k across time, $\sigma(\alpha) = \frac{1}{1+\exp(-\alpha)}$ is the sigmoid function, and $\alpha_{k,m}$ are learnable weights. The quantity $\sigma(\alpha_{k,m}) \in (0, 1)$ determines the point at which the approximation $\tilde{Q}_{\mathcal{X}_k}$ to the empirical quantile function of the set $\mathcal{X}_k$ is sampled. As $\sigma(\alpha_{k,m})$ is varied from near zero to near one, $y_{k,m}$ will change continuously from the minimum to the maximum of the values in $\mathcal{X}_k$, following the piecewise linear function in Figure 2. Due to the piecewise linear approximation, gradients of the weights $\alpha_{k,m}$ with respect to the network loss are nonzero and the layer can be trained end-to-end using standard stochastic gradient methods.

The quantile layer is easily implemented in deep learning frameworks by sorting the incoming activations for each filter k, linearly interpolating, and returning the linearly interpolated values at the points prescribed by weights $\alpha_{k,1}, ..., \alpha_{k,M}$. The output of the layer is a discrete approximation to the empirical quantile function of the activations of filter k. The learnable weights determine at which part of the cumulative distribution function the approximation is focused. For example, sampling points can be spaced uniformly across the spectrum of values or concentrate on those values that are near the maximum or minimum.

4 Model Architectures

We treat user identification from gaze movement patterns as a sequence classification problem. The input is a sequence of two-dimensional gaze positions, separately recorded for the left and right eye, and sampled regularly over time. The data we work with additionally contains a scalar measurement of the pupil dilation for the left and the right eye at each point in time. We concatenate the gaze positions and pupil dilations to form a sequence of shape $T \times 6$, where the sequence length T is typically different for each input.

We study 1D-convolutional neural networks to classify gaze movement sequences, using two different architectures. The first architecture stacks 1D-convolution layers to extract local features from the sequence without reducing the temporal dimension by intermediate pooling layers; the temporal dimension is then aggregated out in a statistical aggregation layer before classification is performed. The second architecture reduces the temporal dimension with intermediate pooling layers to capture more large-scale temporal patterns before performing aggregation. Both architectures are 17 layers deep (not including pooling or aggregation layers) and are shown in Table 1. As aggregation layer, we study the quantile layer introduced in Section 3, global maximum pooling, global average pooling, and the histogram layers proposed by Wang et al. [30] and Sedighi and Fridrich [27]. More details about baselines are given in Section 5.

All convolution layers are followed by a nonlinear activation function. We use parameterized ReLU activations [8], a generalization of leaky ReLUs, of the form

$$s(y) = \begin{cases} y & \text{if } y > 0 \\ (1 - \beta_j)y & \text{if } y \leq 0. \end{cases} \tag{11}$$

Table 1. Network architectures without (left) and with (right) intermediate pooling layers. T denotes the sequence length. All convolution layers use stride one, the pooling layers use stride two. Both architectures use dropout with parameter 0.5 before the fully connected layer. As aggregation layer we study the quantile layer, global maximum or average pooling, and the histogram layers by Wang et al. [30] and Sedighi and Fridrich [27]. Output shape M and parameters vary across aggregation layers.

Architecture Without Intermediate Pooling		Architecture With Intermediate Pooling		
Layer	Output Size	Layer	Output Size	Parameters
input	$T \times 6$	input	$T \times 6$	0
$\left[\text{conv } 3 \times 1 - 16\right] \times 4$	$T \times 16$	$\left[\text{conv } 3 \times 1 - 16\right] \times 4$	$T \times 16$	2660
-	-	pool 2×1	$T/2 \times 16$	0
$\left[\text{conv } 3 \times 1 - 32\right] \times 4$	$T \times 32$	$\left[\text{conv } 3 \times 1 - 32\right] \times 4$	$T/2 \times 32$	10884
-	-	pool 2×1	$T/4 \times 32$	0
$\left[\text{conv } 3 \times 1 - 64\right] \times 4$	$T \times 64$	$\left[\text{conv } 3 \times 1 - 64\right] \times 4$	$T/4 \times 64$	43268
-	-	pool 2×1	$T/8 \times 64$	0
$\left[\text{conv } 3 \times 1 - 128\right] \times 4$	$T \times 128$	$\left[\text{conv } 3 \times 1 - 128\right] \times 4$	$T/8 \times 128$	172548
aggregation	$128 \times M$	aggregation	$128 \times M$	variable
fully connected	210	fully connected	210	$27090 \cdot M$

where β_j is a layer-specific parameter and j is the layer index. The parameters β_j are fitted during training and regularized towards zero, such that the slope of the activation below zero does not become too small. The rationale for using this activation is that we want to preserve as much information as possible about the distribution of the responses of the convolution filters, so that this information can later be exploited in the statistical aggregation layer. In contrast, regular ReLU activations discard much information by not distinguishing between any activation values that fall below zero.

As an alternative to the 1D-convolutional architectures shown in Table 1, we also study a recurrent neural network architecture. We choose gated recurrent units (GRU, [3]) as the recurrent unit, because we found architectures based on GRUs to be faster and more robust to train and these architectures have been shown to yield very similar predictive performance [4] as architectures based on LSTM units [10]. We study a sequence classification architecture in which the input layer is followed by two layers of gated recurrent units, and the state vector of the last GRU in the second layer is fed into a dense layer that predicts the class label. The first layer of GRUs contains 64 units and the second layer 128 units. We employ dropout with dropout parameter 0.5 before the dense layer.

5 Empirical Study

In this section, we empirically study how accurately subjects can be distinguished based on observed gaze patterns. We evaluate different neural network architectures and aggregation layers, and compare with existing probabilistic and feature-based models for eye gaze biometrics.

5.1 Experimental Setup

Data The *Dynamic Images and Eye Movements* (DIEM) project is a large-scale data collection effort in which gaze movements of subjects have been recorded while viewing non-text visual input [21]. The DIEM data set contains gaze movement observations of 223 subjects on 85 short video sequences that contain a variety of visual material, such as recordings of street scenes, documentary videos, movie excerpts, recordings of sport matches, or television advertisements. Subjects in the data set have viewed between 6 and 26 videos. We restrict ourselves to those subjects which have viewed at least 25 videos, which leaves 210 of the 223 subjects in the data. The average length of a video sequence is 95 seconds. The entire data set contains 5381 gaze movement sequences.

Gaze movements have been recorded with an SR Research Eyelink 2000 eye tracker. While the original temporal resolution of the eye tracker is 1000 Hz, in the DIEM data set gaze movements are sampled down to a temporal resolution of 30 Hz [21]. This is a lower resolution than used in most other studies; for example, Abdelwahab et al. [1] use 500 Hz, while studies by Holland and Komogortsev [11, 12] use either 1000 Hz or 75 Hz data. At each of the 30 time points per second, the two-dimensional gaze position and a scalar measurement of the pupil dilation is available for the left and the right eye, which we concatenate to form a six-dimensional input.

Problem Setup We treat the problem of identifying individuals in the DIEM data set based on their gaze patterns as a 210-class classification problem. A training instance is a sequence of gaze movements (of one individual on one video), annotated with the individual's identity as the class label. We split the entire set of 5381 gaze movement sequences into a training set (2734 sequences), a validation set (537 sequences), and a test set (2110 sequences). The split is constructed by splitting the 84 videos into 50% (42) training videos, 10% (8) validation videos, and 40% (34) test videos, and including the gaze movement observations of all individuals on the training, validation, and test videos in the respective set of sequences. This ensures that predictions are evaluated on novel visual input not seen in the training data. At test time, the task is to infer the unknown identity of an individual after observing gaze patterns of that individual on N video sequences drawn at random from all videos in the test set viewed by that individual, where N is varied from one to five. Applying a learned model to each of the N sequences yields predictive class probabilities $p_{i,j}$ for $1 \leq i \leq N$ and $1 \leq j \leq 210$. The most likely identity is then inferred by

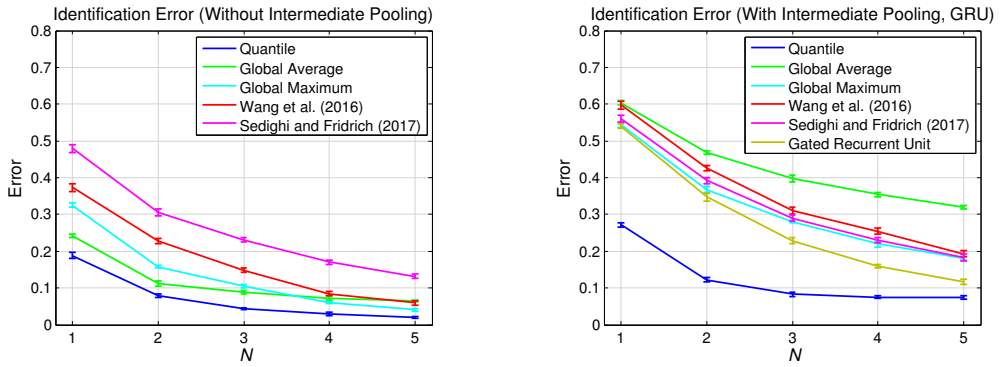

Fig. 3. Identification error for convolutional neural network architectures without intermediate pooling (left), with intermediate pooling (right) and for the recurrent neural network architecture (right) as a function of the number of test videos N on which a user is observed. Error bars indicate the standard error.

$\arg\max_j \prod_{i=1}^{N} p_{i,j}$ and compared to the true identity. We measure *identification error*, defined as the fraction of experiments in which the inferred identity is not equal to the true identity of the individual. Results are averaged over the 210 individuals and 10 random draws of test videos for each individual.

Methods under Study We study the deep neural network architectures with and without intermediate pooling layers shown in Table 1 in combination with different aggregation layers: the quantile layer as described in Section 3 (*Quantile*), global maximum or average pooling (*Global Maximum*, *Global Average*), and the histogram layers proposed by *Wang et al. [30]* and *Sedighi and Fridrich [27]*. The input to the histogram layers is identical to the input of the quantile layer, namely a matrix $\mathbf{Z} \in \mathbb{R}^{T \times K}$ of activations of the preceding convolution layer. The layers approximate the distribution of values per filter k in $\mathbf{Z}$ by a histogram with M bins, where bin centers and bin widths are learnable parameters. The output is a matrix $\mathbf{Y} \in \mathbb{R}^{K \times M}$; an element $y_{k,m}$ of the output computes the fraction of values of filter k that fall into bin m. The two histogram baselines differ in how they smoothen the sharp edges in traditional histogram functions in order to enable gradient flow: using linear approximations [30] or Gaussian kernels [27]. For the models with quantile and histogram layers, the hyperparameter M is optimized on the validation set on a grid $M \in \{4, 8, 16, 32\}$, yielding $M = 8$ for both histogram-based models and $M = 16$ for the quantile-based model. We use the Adam optimizer [14] with initial learning rate 0.0001 and train all models for 2000 epochs. For histogram-based models, optimization failed with the default initial learning rate of 0.0001. We instead use an initial learning rate of 0.00001, with which optimization succeeded. The batch size is one in all experiments.

We also study the recurrent neural network architecture with two hidden layers of gated recurrent units as discussed in Section 4. It is trained with the Adam optimizer for 2000 episodes, using an initial learning rate of 0.001.

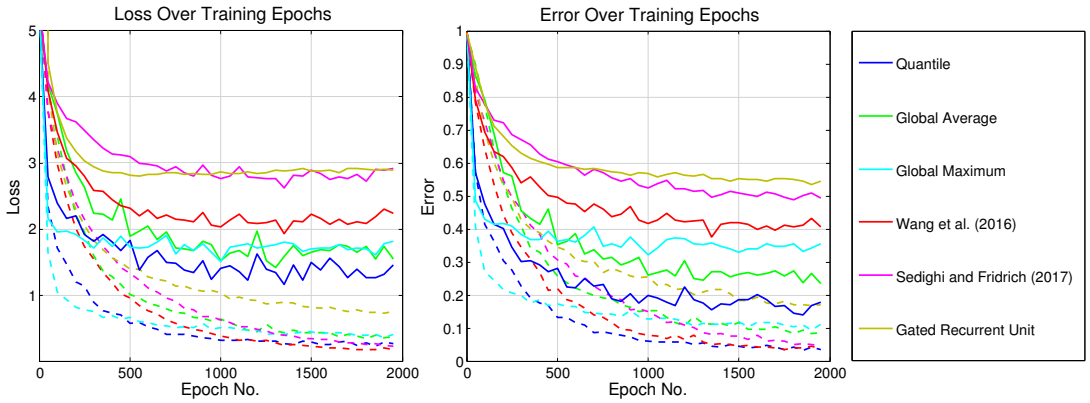

Fig. 4. Identification error (left) and loss (right) for convolutional network architectures without intermediate pooling and recurrent neural network as a function of the epoch number during training. Dashed curves denote training error and loss while solid curves denote test error and loss.

As further baselines, we study the probabilistic approaches by Kinnunen et al. [15], Landwehr et al. [17], and Abdelwahab et al. [1], which respectively employ Gaussian mixture models, parametric models based on the Gamma family, and semiparametric models based on Gaussian processes in order to characterize distributions over gaze patterns. The model of Kinnunen et al. can be directly applied in our domain. We tune the number of histogram bins, window size, and number of mixture components on the validation data. The models of Landwehr et al. [17] and Abdelwahab et al. [1] were designed for gaze movements during reading; they are therefore not directly applicable. We adapt these models of to our non-text domain as follows. Both models characterize individual gaze patterns by separately fitting the distribution of saccade amplitudes and fixation durations for different so-called saccade types: *regression*, *refixation*, *next word movement*, and *forward skip*. The saccade types relate the gaze movement to the structure of the text being read. We instead separately fit distributions for saccade types *up*, *down*, *left*, *right*, which indicate the predominant direction of the gaze movement. The DIEM data contains saccade and fixation annotations; we can thus preprocess the data into sequences of saccades and fixations as needed for an empirical comparison with these models. Another recently published probabilistic model is that of Makowski et al. [20]. This model is more difficult to adapt because it is built around lexical features of the text being read; without lexical features it was empirically found to be no more accurate than the model by Abdelwahab et al. [20]. We therefore exclude it from the empirical study.

We finally compare against the feature-based methods of Holland and Komogortsev [12] and Rigas et al. [26]. Both of these methods follow the same general approach, only using different sets of features. We use the variant that employs two-sample Kolmogorov-Smirnov test for the matching module and weighted mean as the fusion method, since results reported in the paper were best for these variants on low-resolution data [12].

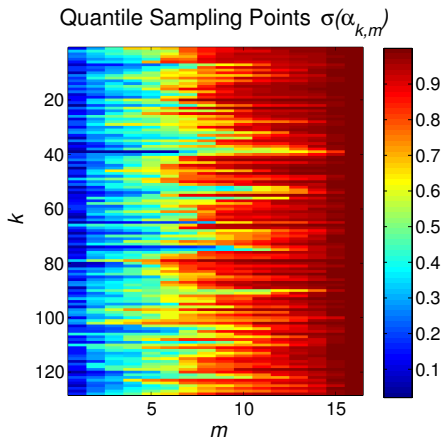

Fig. 5. Learned quantile sampling points $\sigma(\alpha_{k,m})$ as defined by Equation 10.

5.2 Results

Figure 3 shows error rates for identifying individuals in the DIEM data set for different neural network architectures, including the recurrent neural network, as a function of the number N of test videos on which gaze patterns of the unknown individual are observed. We observe that architectures without intermediate pooling layers have lower error rates. This is in line with the assumption that local, short-term gaze patterns are most informative for identification: the larger receptive fields of neurons in architectures with intermediate pooling do not appear to be advantageous. We will therefore focus on architectures without intermediate pooling in the remaining discussion. Architectures based on gated recurrent units are also focused on fitting relatively long-term temporal patterns in data; the recurrent architecture we study performs slightly better than convolutional architectures with intermediate pooling but worse than convolutional architectures without intermediate pooling. Employing quantile layers for statistical aggregation outperforms global maximum or average pooling, indicating that retaining more information about the distribution of filter activations is informative for identification. Surprisingly, architectures based on the histogram layers proposed by Wang et al. [30] and Sedighi and Fridrich [27] do not consistently improve over the global pooling methods.

Figure 4 shows error rates and losses for architectures without intermediate pooling layers on the training and test data as a function of the epoch number during training. We observe that architectures with quantile and histogram layers both achieve lower training error than architectures with global maximum or average pooling, but only for the quantile-based model this translates into lower error on the test data. Figure 4 thus does not suggest that there are any problems with fitting the histogram-based models using our training protocol; manual inspection of the learned histogram bins also showed reasonable bin centers and widths. Rather, results seem to indicate that characterizing distributions in terms of quantiles – which is closer to standard average or maximum pooling operations – generalizes better than characterizing distributions by histograms.

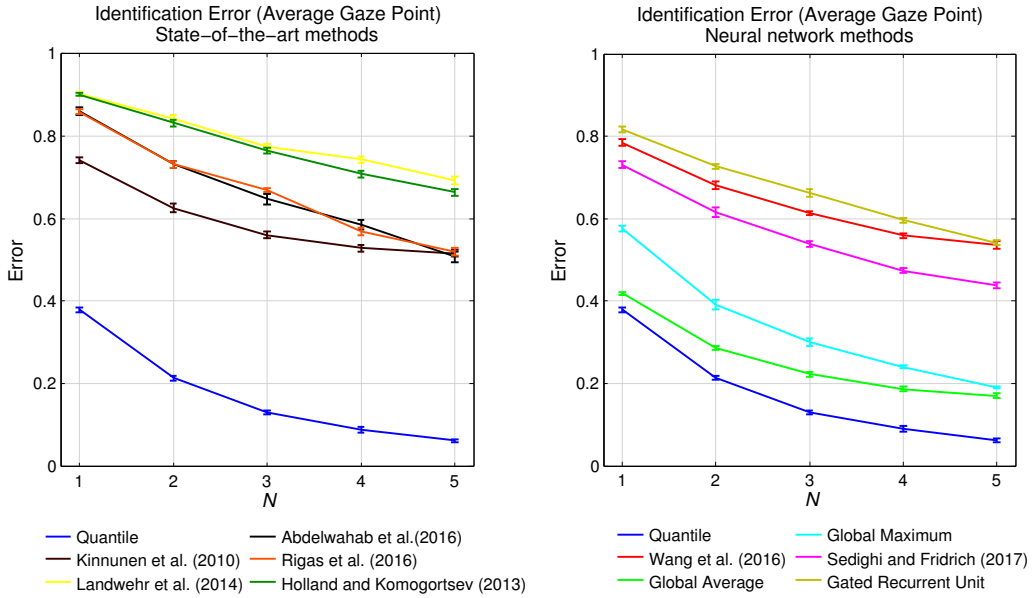

Fig. 6. Identification error as a function of the number of test videos N on which a user is observed, using average gaze point only. Error bars indicate the standard error.

Figure 5 shows learned values for the quantile sampling points $\sigma(\alpha_{k,m})$ (see Equation 10). We observe that sampling points adapt to each filter, and outputs $y_{k,m}$ of the quantile layer focus more on values close to the maximum ($\sigma(\alpha_{k,m})$ near one) than the minimum ($\sigma(\alpha_{k,m})$ near zero).

We finally compare against probabilistic and feature-based baselines from the literature, specifically the models of Kinnunen et al. [15], Landwehr et al. [17], Abdelwahab et al. [1], Holland and Komogortsev [12] and Rigas et al. [26]. These models only use the gaze position averaged over the left and right eye, and do not use pupil dilation. We also study our models in this setting, using only the average gaze position as input in the neural networks. Figure 6 shows identification error as a function of the number of test videos for this setting. We observe that identification errors are generally higher than in the setting where separate gaze positions and pupil dilations are available. Moreover, the best neural networks outperform the probabilistic and feature-based models by a large margin. This may partially be explained by the fact that the probabilistic models were originally developed for text reading, and for data with a much higher temporal resolution (500 Hz versus 30 Hz in our study). The quantile-based model again performs best among the neural network architectures studied.

6 Conclusions

We have studied deep neural networks for unobtrusive biometric identification based on gaze patterns observed on non-text visual input. Differences in the distribution of local, short-term gaze patterns are most informative for distinguishing between individuals. To characterize these distributions, we introduced

the quantile layer, a learnable statistical aggregation layer that approximates the empirical quantile function of the activations of a preceding stack of 1D-convolution layers. In contrast to existing learnable statistical aggregation layers that approximate the distribution of filter activations by a histogram, the quantile layer naturally generalizes standard global pooling layers. From our empirical study we can conclude that neural networks with quantile layers outperform networks with global average or maximum pooling, as well as networks that use histogram layers. In our domain, deep neural networks also outperform probabilistic and feature-based models from the literature by a wide margin.

Acknowledgments

This work was partially funded by the German Research Foundation under grant LA3270/1-1.

References

1. Abdelwahab, A., Kliegl, R., Landwehr, N.: A semiparametric model for Bayesian reader identification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-2016). Austin, TX (2016)
2. Bednarik, R., Kinnunen, T., Mihaila, A., Fränti, P.: Eye-movements as a biometric. In: Proceedings of the 14th Scandinavian Conference on Image Analysis (2005)
3. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078 (2014)
4. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555 (2014)
5. Couture, H.D., Marron, J., Perou, C.M., Troester, M.A., Niethammer, M.: Multiple Instance Learning for Heterogeneous Images: Training a CNN for Histopathology. In: Proceedings of the 2018 International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 254–262 (2018)
6. Dufour, J.M.: Distribution and quantile functions. Tech. rep., McGill University, Montreal, Canada (1995)
7. George, A., Routray, A.: A score level fusion method for eye movement biometrics. Pattern Recognition Letters **82**(2), 207–215 (2016)
8. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
9. Henderson, J.M.: Human gaze control during real-world scene perception. Trends in cognitive sciences **7**(11), 498–504 (2003)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
11. Holland, C., Komogortsev, O.V.: Biometric identification via eye movement scanpaths in reading. In: Proceedings of the 2011 International Joint Conference on Biometrics (2012)
12. Holland, C.D., Komogortsev, O.V.: Complex eye movement pattern biometrics: Analyzing fixations and saccades. In: 2013 International conference on biometrics (ICB). pp. 1–8. IEEE (2013)

13. Kasprowski, P., Ober, J.: Eye movements in biometrics. In: Proceedings of the 2004 International Biometric Authentication Workshop (2004)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
15. Kinnunen, T., Sedlak, F., Bednarik, R.: Towards task-independent person authentication using eye movement signals. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications. pp. 187–190. ACM (2010)
16. Komogortsev, O.V., Jayarathna, S., Aragon, C.R., Mahmoud, M.: Biometric identification via an oculomotor plant mathematical model. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (2010)
17. Landwehr, N., Arzt, S., Scheffer, T., Kliegl, R.: A model of individual differences in gaze control during reading. In: Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing (2014)
18. Liang, Z., Tan, F., Chi, Z.: Video-based biometric identification using eye tracking technique. In: Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on. pp. 728–733. IEEE (2012)
19. Liversedge, S.P., Findlay, J.M.: Saccadic eye movements and cognition. Trends in cognitive sciences **4**(1), 6–14 (2000)
20. Makowski, S., Jäger, L., Abdelwahab, A., Landwehr, N., Scheffer, T.: A discriminative model for identifying readers and assessing text comprehension from eye movements. In: Proceedings of the 29th European Conference on Machine Learning (ECML-2018). Dublin, Ireland (2018)
21. Mital, P.K., Smith, T.J., Hill, R.L., Henderson, J.M.: Clustering of gaze during dynamic scene viewing is predicted by motion. Cognitive Computation **3**(1), 5–24 (2011)
22. Poynter, W., Barber, M., Inman, J., Wiggins, C.: Individuals exhibit idiosyncratic eye-movement behavior profiles across tasks. Vision Research **89**, 32 – 38 (2013)
23. Rayner, K., Li, X., Williams, C.C., Cave, K.R., Well, A.D.: Eye movements during information processing tasks: Individual differences and cultural effects. Vision Research **47**(21), 2714 – 2726 (2007)
24. Resnick, S.I.: Extreme values, regular variation and point processes. Springer (2013)
25. Rigas, I., Economou, G., Fotopoulos, S.: Biometric identification based on the eye movements and graph matching techniques. Pattern Recognition Letters **33**(6) (2012)
26. Rigas, I., Komogortsev, O., Shadmehr, R.: Biometric recognition via eye movements: Saccadic vigor and acceleration cues. ACM Transaction on Applied Perception **13**(2), 1–21 (2016)
27. Sedighi, V., Fridrich, J.: Histogram layer, moving convolutional neural networks towards feature-based steganalysis. Electronic Imaging **2017**(7), 50–55 (2017)
28. Van der Vaart, A.W.: Asymptotic statistics, vol. 3. Cambridge university press (2000)
29. Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. arXiv:1609.03499 (2016)
30. Wang, Z., Li, H., Ouyang, W., Wang, X.: Learnable histogram: Statistical context features for deep neural networks. In: European Conference on Computer Vision. pp. 246–262. Springer (2016)
31. Zhang, Y., Juhola, M.: On biometric verification of a user by means of eye movement data mining. In: Proceedings of the 2nd International Conference on Advances in Information Mining and Management (2012)

Deep Distributional Sequence Embeddings Based on a Wasserstein Loss

Ahmed Abdelwahab · Niels Landwehr

Received: date / Accepted: date

Abstract Deep metric learning employs deep neural networks to embed instances into a metric space such that distances between instances of the same class are small and distances between instances from different classes are large. In most existing deep metric learning techniques, the embedding of an instance is given by a feature vector produced by a deep neural network and Euclidean distance or cosine similarity defines distances between these vectors. In this paper, we study deep distributional embeddings of sequences, where the embedding of a sequence is given by the distribution of learned deep features across the sequence. This has the advantage of capturing statistical information about the distribution of patterns within the sequence in the embedding. When embeddings are distributions rather than vectors, measuring distances between embeddings involves comparing their respective distributions. We propose a distance metric based on Wasserstein distances between the distributions and a corresponding loss function for metric learning, which leads to a novel end-to-end trainable embedding model. We empirically observe that distributional embeddings outperform standard vector embeddings and that training with the proposed Wasserstein metric outperforms training with other distance functions.

1 Introduction

Metric learning is concerned with learning a representation or *embedding* in which distances between instances of the same class are small and distances

A. Abdelwahab E-mail: AAbdelwahab@atb-potsdam.de
Leibniz Institute of Agricultural Engineering and Bioeconomy e.V., Potsdam, Germany

N. Landwehr E-mail: Landwehr@cs.uni-potsdam.de
University of Potsdam, Department of Computer Science, Potsdam, Germany
Leibniz Institute of Agricultural Engineering and Bioeconomy e.V., Potsdam, Germany

between instances of different classes are large. Deep metric learning approaches, in which the learned embedding is given by a deep neural network, have achieved state-of-the-art results in many tasks, including face verification and recognition (Schroff et al, 2015), fine-grained image classification (Reed et al, 2016), zero-shot classification (Bucher et al, 2016), speech-to-text problems (Gibiansky et al, 2017), and speaker identification (Li et al, 2017). An advantage of metric learning is that the resulting representation directly generalizes to unseen classes, so the model does not need to be retrained every time a new class is introduced. This is, for example, a typical requirement in biometric applications, where it should be possible to register new subjects without retraining a model. Biometric systems also have to handle imposters, that is, subjects who are not registered in the database, which is not straightforward in standard classification settings.

In this paper, we study deep metric learning for sequence data, with a specific focus on biometric problems. Building on earlier work on *quantile layers* (Abdelwahab and Landwehr, 2019), we specifically study how the distribution of learned deep features across a sequence can be represented in the learned embedding. Quantile layers are statistical aggregation layers that characterize the distribution of patterns within a sequence by approximating the quantile function of the activations of the learned filters across the sequence. Characterizing this distribution has been shown to be advantageous for biometric identification based on eye movement patterns (Abdelwahab and Landwehr, 2019). The main contribution of this paper is to develop a deep metric learning approach for distributional embeddings based on quantile layers. Quantile layers return an estimate of the distribution of values for each learned filter across the sequence. Instead of a fixed-length vector representation of an instance, in our approach, the embedding of an instance is given by these sets of distributions. When embeddings are distributions rather than simple vectors, measuring distances between the embeddings involves comparing their respective distributions. We propose a distance metric in the embedding space that is based on Wasserstein distances between the respective distributions. Compared to other distance functions such as Kulback-Leibler or Jensen-Shannon divergence, the advantage of using Wasserstein distance is that it takes into account the metric on the space in which the random variable of interest is defined. In our case, this means that distributions in which similar magnitudes of filter activations receive similar amounts of probability mass will be considered close. We show how such embeddings can be trained end-to-end on labeled training data using metric learning techniques.

Empirically, we study the proposed approach in biometric identification problems involving eye movement, accelerometer, and EEG data. Empirical results show that the proposed distributional sequence embeddings outperform standard vector embeddings and that training with the Wasserstein metric outperforms training with other distance functions.

The rest of the paper is organized as follows. Section 2 discusses related work. In Section 3 we review quantile layers and develop a distributional embedding architecture based on these layers. Section 4 introduces a Wasserstein-

based distance metric for the proposed embedding model and from this derives a novel loss function for metric learning. In Section 5 we empirically study the proposed method and baselines.

2 Related work

Our work is motivated by the goal of capturing information about the distribution of patterns within a sequence in its embedding, where the patterns are defined in terms of learned features of a deep neural network. It is related to other work in deep learning that aims to capture distributions of learned features using statistical aggregation layers. Wang et al (2016) proposed end-to-end learnable *histogram layers* that approximate the distribution of learned features by a histogram. Their work uses linear approximations to smoothen the sharp edges in a traditional histogram function and enable gradient flow. Sedighi and Fridrich (2017) proposed a similar histogram-based aggregation layer, but use Gaussian kernels as a soft, differentiable approximation to histogram bins. Abdelwahab and Landwehr (2019) introduced *quantile layers* to capture the distribution of learned features based on an approximation of the quantile function, and empirically showed that this outperforms aggregation using histograms. The contribution of our paper is to exploit quantile layers in metric learning, by defining distributional embeddings based on approximations of quantile functions and deriving loss functions for metric learning based on comparing the resulting distributions.

There is a large body of work on deep metric learning that studies different network architectures and loss functions. For example, Hadsell et al (2006) introduced a loss for a siamese network architecture that is based on all possible pairs of instances in the training data, and its objective is to minimize distances between positive pairs (same class) while maximizing the distances between negative pairs (different classes). More recently, Schroff et al (2015) introduced the triplet loss, with links positive and negative pairs by an anchor instance. This idea has later been extended by Oh Song et al (2016) and Sohn (2016) by providing several negative pairs linked to one positive pair to the loss function. The loss function introduced by Sohn (2016) has shown superior performance in several studies (Sohn, 2016; Wu et al, 2017; Yuan et al, 2017). Our method builds on these established deep metric learning techniques, but extends them by replacing vector embeddings with distributional embeddings, which requires corresponding changes in distance calculations and the loss function.

Distributional embeddings have also been studied in natural language processing in the context of word embeddings. Traditional word embedding models such as *word2vec* represent words as vectors in a metric space such that semantically similar words are mapped to similar vectors (Mikolov et al, 2013). Vilnis and McCallum (2015) extend this idea by mapping each word to a Gaussian distribution (with diagonal covariance), which naturally characterizes uncertainty about the embedding. Athiwaratkun and Wilson (2017) further extend this model by replacing the Gaussian distribution with a mixture of Gaus-

sians, where the multimodal mixture can capture multiple meanings of the same word. The motivation for these distributional embeddings is somewhat different from our motivation in this paper: while the distribution in our model results from the inner structure of the instance being mapped (distribution of patterns within a sequence), the distribution in the model by Vilnis and McCallum (2015) captures remaining uncertainty and is inferred during training. Another difference in the work by Vilnis and McCallum (2015) is that their model is trained in an unsupervised fashion, while we study supervised metric learning. An approach similar to that of Vilnis and McCallum (2015) has also been taken by Bojchevski and Günnemann (2018) in order to map nodes of an attributed graph onto Gaussian distributions that function as an embedding representation. This is again an unsupervised approach, and specific to the task of node embedding.

More generally, deep metric learning models have been recently used in different application domains featuring sequential data, including natural language processing (Mueller and Thyagarajan, 2016; Neculoiu et al, 2016), computer vision (McLaughlin et al, 2016; Wu et al, 2018) and speaker identification (Li et al, 2017; Chung et al, 2018), but these approaches are based on vector embeddings rather than distributional embeddings.

3 Quantile Layers and Distributional Sequence Embeddings

This section reviews *quantile layers* as introduced by Abdelwahab and Landwehr (2019) and discusses how they can be used to define distributional embeddings of variable-length sequences.

In this paper, we focus on variable-length sequences and deep convolutional neural network architectures that produce embeddings of such sequences. Typically, network architectures for such sequences would employ stacked convolution layers to extract informative features from the sequence, and in the last layer use some form of global pooling to transform the remaining variable-length representation into a fixed-length vector representation. Global pooling achieves this transformation by performing a simple aggregate operation such as taking the maximum or average over the filter activations across the sequence. This has the potential disadvantage that most information about the distribution of the filter activations is lost, which might be informative for the task at hand. In contrast, quantile layers aim to preserve as much information as possible about the distribution of filter activations along the sequence by approximating the quantile function of this distribution. Earlier work has shown that this information can be informative for sequence classification, substantially increasing predictive accuracy (Abdelwahab and Landwehr, 2019).

In this paper, we use quantile layers for defining distributional embeddings of sequences. We assume that instances are given by variable-length sequences of the form $\mathbf{s} = (\mathbf{x}_1, ..., \mathbf{x}_T)$ where $\mathbf{x}_t \in \mathbb{R}^D$ is a vector of attributes that describes the sequence element at position t. We denote the space of all such sequences with D attributes by $\mathcal{S}_D = \bigcup_{T=1}^{\infty} \mathbb{R}^{T \times D}$. When a sequence is pro-

cessed by a convolutional deep neural network architecture Γ, which we take to be the network without any final global aggregation layers, the result is a variable-length representation of the instance over K filters. We denote this mapping by $\Gamma : \mathcal{S}_D \rightarrow \mathcal{S}_K$. Details of the deep convolutional architectures we employ are given in Section 5. For $\mathbf{s} \in \mathcal{S}_D$ and $k \in \{1, ..., K\}$ we will use $\Gamma_k(\mathbf{s})$ to denote the variable-length sequence of activations of filter k produced by the network for sequence $\mathbf{s}$.

As in Abdelwahab and Landwehr (2019) we use quantile functions in order to characterize the distribution of filter activations across the sequence $\Gamma_k(\mathbf{s})$. Let $x \in \mathbb{R}$ be a real-valued random variable, let $p(x)$ denote its density and $F(x)$ its cumulative distribution function. The quantile function for x is defined by

$$Q(r) = \inf\{x \in \mathbb{R} : F(x) \geq r\}$$

where inf denotes the infimum. If F is continuous and strictly monotonically increasing, Q is simply the inverse of F. Let $\mathcal{X} = \{x_1, ..., x_N\}$ be a sample of the random variable x, that is, $x_n \sim p(x)$ for $n \in \{1, ..., N\}$. The empirical quantile function $\hat{Q}_{\mathcal{X}} : (0, 1] \rightarrow \mathbb{R}$ is a non-parametric estimator of the quantile function Q. It is defined by

$$\hat{Q}_{\mathcal{X}}(r) = \inf\{x \in \mathbb{R} : r \leq \hat{F}_{\mathcal{X}}(x)\} \tag{1}$$

where $\hat{F}_{\mathcal{X}}(x) = \frac{1}{N} \sum_{i=1}^{N} I(x_i \leq x)$ is the empirical cumulative distribution function and $I(x_i \leq x) \in \{0, 1\}$ is an indicator. $\hat{Q}_{\mathcal{X}}(r)$ is a piecewise constant function that is essentially obtained by sorting the samples in $\mathcal{X}$. More formally, let π be a permutation that sorts the x_i, that is, $x_{\pi(i)} \leq x_{\pi(i+1)}$ for $1 \leq i \leq N - 1$. Then $\hat{Q}_{\mathcal{X}}(r) = x_{\pi(\lceil rN \rceil)}$, where $\lceil x \rceil$ denotes the smallest integer larger or equal to x. The empirical quantile function $\hat{Q}_{\mathcal{X}}$ faithfully approximates the quantile function Q in the sense that $|\hat{Q}_{\mathcal{X}}(r) - Q(r)|$ converges almost surely to zero if $N \rightarrow \infty$ and Q is continuous at r (Resnick, 2013).

To enable gradient flow in end-to-end learning, we will work with a piecewise linear interpolation of the piecewise constant function $\hat{Q}_{\mathcal{X}}(r)$. For $i \in \{1, ..., N\}$ and $r \in [\frac{n-1}{N}, \frac{n}{N}]$ we can define a linear approximation by

$$\tilde{Q}_{\mathcal{X}}(r) = N(x_{\pi(n+1)} - x_{\pi(n)})r + nx_{\pi(n)} + (1 - n)x_{\pi(n+1)} \quad \left(r \in \left[\frac{n-1}{N}, \frac{n}{N}\right]\right)$$

where we define $x_{\pi(N+1)} = x_{\pi(N)}$ to handle the right interval border. Combining the linear approximations over the different n, we obtain for $r \in [0, 1]$ the piecewise linear approximation

$$\tilde{Q}_{\mathcal{X}}(r) = \sum_{n=1}^{N} \tilde{\delta}(r, n) \left(N(x_{\pi(n+1)} - x_{\pi(n)})r + nx_{\pi(n)} + (1 - n)x_{\pi(n+1)}\right)$$

where $\tilde{\delta}(r, n)$ is an indicator function that is defined as one if $r \in [\frac{n-1}{N}, \frac{n}{N}]$ and zero otherwise. The piecewise linear approximation $\tilde{Q}_{\mathcal{X}}(r)$ of the quantile function depends on the sample size N, because there are N linear segments. To arrive at an approximation of the quantile function that is independent of the number of samples, we define a further piecewise linear

approximation of $\tilde{Q}_{\mathcal{X}}(r)$ using M sampling points $\sigma(\alpha_1), ..., \sigma(\alpha_M)$, where $\sigma(\alpha) = (1 + exp(-\alpha))^{-1}$ is the sigmoid function and $\alpha_i \in \mathbb{R}$ are parameters with $\alpha_i \leq \alpha_{i+1}$. Formally, we define

$$\bar{Q}_{\mathcal{X}}(r) = \sum_{i=0}^{M} \bar{\delta}(r, i)(a_{\mathcal{X},i} r + b_{\mathcal{X},i}) \tag{2}$$

where

$$a_{\mathcal{X},i} = \frac{\tilde{Q}_{\mathcal{X}}(\sigma(\alpha_{i+1})) - \tilde{Q}_{\mathcal{X}}(\sigma(\alpha_i))}{\sigma(\alpha_{i+1}) - \sigma(\alpha_i)} \tag{3}$$

$$b_{\mathcal{X},i} = \tilde{Q}_{\mathcal{X}}(\sigma(\alpha_i)) - \sigma(\alpha_i) \frac{\tilde{Q}_{\mathcal{X}}(\sigma(\alpha_{i+1})) - \tilde{Q}_{\mathcal{X}}(\sigma(\alpha_i))}{\sigma(\alpha_{i+1}) - \sigma(\alpha_i)}, \tag{4}$$

$\bar{\delta}(r, i)$ is an indicator function that is one if $r \in [\sigma(\alpha_i), \sigma(\alpha_{i+1})]$ and zero otherwise, and we have introduced $\alpha_0 = -\infty$ and $\alpha_{M+1} = \infty$ to handle border cases. The function $\bar{Q}_{\mathcal{X}}(r)$ provides a piecewise linear approximation of the quantile function using $M+1$ line segments, independently of the sample size N. The parameters α_i are learnable model parameters in the deep neural network architectures that we study in Section 5.

We are now ready to define the distributional embedding for an instance, which is obtained by passing the instance through the neural network Γ and for each filter in the output of Γ approximating the quantile function of the filter activations by the piecewise linear function $\bar{Q}$.

Definition 1 (Distributional embedding of sequence) Let $\mathbf{s} \in \mathcal{S}_D$ and let Γ denote a convolutional neural network structure. The distributional embedding of sequence $\mathbf{s}$ is given by the vector of piecewise linear functions

$$\Psi_{\Gamma}(\mathbf{s}) = \left(\bar{Q}_{\Gamma_1(\mathbf{s})}, ..., \bar{Q}_{\Gamma_K(\mathbf{s})}\right) \tag{5}$$

where $\bar{Q}_{\Gamma_k(\mathbf{s})}$ is defined by Equation 2 using $\mathcal{X} = \Gamma_k(\mathbf{s})$. Here, we slightly generalize the notation by identifying the sequence of observations $\Gamma_k(\mathbf{s})$ with the corresponding set of observations.

We note that due to the piecewise linear approximations, gradients can flow through the entire embedding architecture, both to parameters α_m and the weights in the deep neural network structure Γ. This includes the sorting operation, where gradients can be passed through by reordering the gradient backpropagated from the layer above according to the sorting indices π.

4 A Wasserstein Loss for Distributional Embeddings

For training the embedding model, we will use deep metric learning approaches which train model parameters such that instances of the same class are close and instances of different classes are far apart in the embedding space. In order to apply such approaches, a distance metric needs to be defined on the embedding space.

4.1 Distances Between Distributional Embeddings

As discussed in Section 3, in our setting embeddings of instances are given by distributions. Measuring the distance between two embeddings thus means comparing their respective distributions. Different approaches to measure distances between probability distributions have been discussed in the literature. One of the most widely used distance functions between distributions is the Kullback-Leibler divergence. However, this measure is asymmetric and can result in infinite distances, and is therefore not a metric. A metric based on the Kullback-Leibler divergence is the square root of the Jensen-Shannon divergence, which is symmetric, bounded between zero and $\sqrt{log(2)}$, and satisfies the triangle inequality. However, this metric does not yield useful gradients in case the distributions being compared have disjoint support, which in our case would occur if two sequences with non-overlapping ranges of filter values are compared. To illustrate, let q_1 and q_2 denote densities with disjoint support A_1 and A_2, and let $m(x) = \frac{q_1(x)+q_2(x)}{2}$. Then the Jensen-Shannon divergence J of q_1 and q_2 is

$$\begin{aligned} J(q_1, q_2) &= \frac{1}{2}\int_{A_1 \cup A_2} q_1(x) log\left(\frac{q_1(x)}{m(x)}\right)dx + \frac{1}{2}\int_{A_1 \cup A_2} q_2(x) log\left(\frac{q_2(x)}{m(x)}\right)dx \\ &= \frac{1}{2}\int_{A_1} q_1(x) log\left(2\frac{q_1(x)}{q_1(x)}\right)dx + \frac{1}{2}\int_{A_2} q_2(x) log\left(2\frac{q_2(x)}{q_2(x)}\right)dx \\ &= log(2) \end{aligned}$$

independently of the distance between A_1 and A_2, resulting in a gradient of zero.

A different class of distance functions which are increasingly being studied in machine learning (Frogner et al, 2015; Gao and Kleywegt, 2016; Arjovsky et al, 2017) are Wasserstein distances. Wasserstein distances are based on the idea of optimal transport plans. They do not suffer from the zero-gradient problem exhibited by the Jensen-Shannon divergence, because they take into account the metric of the underlying space. They also guarantee continuity under mild assumptions, which is not the case for the Jensen-Shannon divergence as illustrated by Arjovsky et al (2017). In the general case, the p-Wasserstein distance (for $p \in \mathbb{N}$) between two probability measures ρ_1 and ρ_2 over a space $\mathcal{M}$ with metric d can be defined as

$$W_p(\rho_1, \rho_2) = \left(\inf_{\pi \in \mathcal{J}(\rho_1, \rho_2)} \int_{\mathcal{M} \times \mathcal{M}} d(x, y)^p d\pi(x, y)\right)^{\frac{1}{p}} \tag{6}$$

where $\mathcal{J}(\rho_1, \rho_2)$ denotes the set of all joint measures on $\mathcal{M} \times \mathcal{M}$ with marginals ρ_1 and ρ_2. For the purpose of this paper, we are interested in the case of real-valued random variables. If $q_1(x_1)$ and $q_2(x_2)$ are two densities defining distributions over real-valued random variables, $x_i \in \mathbb{R}$, the p-Wasserstein distance between q_1 and q_2 is given by

$$W_p(q_1, q_2) = \left(\inf_{q \in \mathcal{J}(q_1, q_2)} \iint |x_1 - x_2|^p q(x_1, x_2) dx_1 dx_2\right)^{\frac{1}{p}} \tag{7}$$

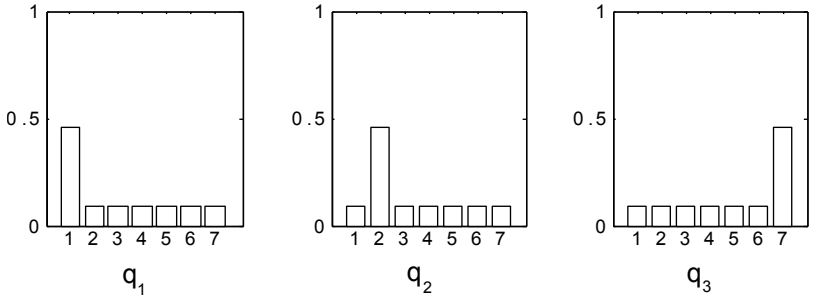

Fig. 1 According to the Wasserstein metric, distributions q_1 and q_2 are closer than q_1 and q_3, while distances would be identical under the Jensen-Shannon measure.

where $\mathcal{J}(q_1, q_2)$ defines the set of all joint distributions over x_1, x_2 which have marginals q_1 and q_2. A joint distribution $q \in \mathcal{J}(q_1, q_2)$ can be seen as a *transport plan*, that is, a way of moving probability mass from density q_1 such that the resulting density is q_2, in the sense that $q(x_1, x_2)$ indicates how much mass is moved from $q_1(x_1)$ to $q_2(x_2)$. The quantity $\iint |x_1 - x_2|^p q(x_1, x_2) dx_1 dx_2$ is the cost of the transport plan, which depends on the amount of probability mass moved, $q(x_1, x_2)$, and the distance by which the mass has been moved, $|x_1 - x_2|^p$. The infimum over the set $\mathcal{J}(q_1, q_2)$ means that the distance between the distributions is given by the optimal transport plan, which intuitively characterizes the minimum changes that need to be made to q_1 in order to transform it into q_2. For $p = 1$ the distance is therefore also called the *Earth Mover Distance*. The advantage of this measure is that it takes into account the metric in the underlying space, as can be seen from Figure 1. Here, q_1 is closer to q_2 than it is to q_3 in the sense that the probability mass needs to be moved less far. Thus, $W_p(q_1, q_2) < W_p(q_1, q_3)$, while the Jensen-Shannon distances between the two pairs of distributions would be identical.

Because Wasserstein distances are defined in terms of optimal transport plans, computing them in general requires solving non-trivial optimization problems. However, for the case of real-valued random variables $x_i \in \mathbb{R}$, there is a simple closed-form solution to the infimum in Equation 7. Let $x_1 \sim q_1$, $x_2 \sim q_2$ with $x_i \in \mathbb{R}$. According to Cambanis et al (1976), the function $K(x_1, x_2) = |x_1 - x_2|^p$ for $p \geq 1$ is quasi-antitone and therefore the infimum of the expectation of this function over the set of all joint distributions, $\inf_{q \in \mathcal{J}(q_1, q_2)} E[K(x_1, x_2)]$, is given by $\int_0^1 K(Q_1(r), Q_2(r)) dr$, where $Q_i(r) = \inf\{t : q_i(x_i \leq t) \geq r\}$ is the quantile function to the density q_i. We can thus rewrite Equation 7 as

$$W_p(q_1, q_2) = \left(\int_0^1 |Q_1(r) - Q_2(r)|^p dr \right)^{\frac{1}{p}}. \tag{8}$$

We now define the distance between two embeddings $\Psi_\Gamma(\mathbf{s})$ and $\Psi_\Gamma(\mathbf{s}')$ as the Wasserstein distance between the approximate representation of the quantile functions in the embedding as defined by Definition 1, summed over the different filters k.

Definition 2 Let $\mathbf{s}, \mathbf{s}' \in \mathcal{S}_D$, let Γ denote a convolutional neural network architecture, and let $\Psi_\Gamma(\mathbf{s})$ and $\Psi_\Gamma(\mathbf{s}')$ denote the distributional embeddings of $\mathbf{s}$, $\mathbf{s}'$ as defined by Definition 1. Then we define the distance between the embeddings as

$$d_p(\Psi_\Gamma(\mathbf{s}), \Psi_\Gamma(\mathbf{s}')) = \sum_{k=1}^{K} \left(\int_0^1 |\bar{Q}_{\Gamma_k(\mathbf{s})}(r) - \bar{Q}_{\Gamma_k(\mathbf{s}')}(r)|^p dr \right)^{\frac{1}{p}} \tag{9}$$

The next proposition gives a closed-form result for computing $d_p(\Psi_\Gamma(\mathbf{s}), \Psi_\Gamma(\mathbf{s}'))$.

Proposition 1 *Let $\mathbf{s}, \mathbf{s}' \in \mathcal{S}_D$, let Γ denote a convolutional neural network architecture, let $\Psi_\Gamma(\mathbf{s})$ and $\Psi_\Gamma(\mathbf{s}')$ denote the distributional embeddings of $\mathbf{s}$, $\mathbf{s}'$, and let $d_p(\Psi_\Gamma(\mathbf{s}), \Psi_\Gamma(\mathbf{s}'))$ denote their distance as defined by Definition 2. Then*

$$d_p(\Psi_\Gamma(\mathbf{s}), \Psi_\Gamma(\mathbf{s}')) = \sum_{k=1}^{K} \Bigg(\sum_{i=0}^{M} \frac{(\bar{a}_{i,k}\sigma(\alpha_{i+1}) + \bar{b}_{i,k})|\bar{b}_{i,k}\sigma(\alpha_{i+1}) + \bar{b}_{i,k}|^p}{\bar{a}_{i,k}(p+1)} - \frac{(\bar{a}_{i,k}\sigma(\alpha_i) + \bar{b}_{i,k})|\bar{a}_{i,k}\sigma(\alpha_i) + \bar{b}_{i,k}|^p}{\bar{a}_{i,k}(p+1)} \Bigg)^{\frac{1}{p}} \tag{10}$$

with

$$\bar{a}_{i,k} = a_{\Gamma_k(\mathbf{s}),i} - a_{\Gamma_k(\mathbf{s}'),i}$$
$$\bar{b}_{i,k} = b_{\Gamma_k(\mathbf{s}),i} - b_{\Gamma_k(\mathbf{s}'),i}$$

where $a_{\mathcal{X},i}$ and $b_{\mathcal{X},i}$ for $\mathcal{X} \in \{\Gamma_k(\mathbf{s}), \Gamma_k(\mathbf{s}')\}$ are defined by Equations 3 and 4, σ is the sigmoid function, and as above we have introduced $\alpha_0 = -\infty$ and $\alpha_{M+1} = \infty$ to handle border cases.

Proof (Proposition 1) Starting from Definition 2 and plugging in $\bar{Q}_{\Gamma_k(\mathbf{s})}$ as defined by Equation 2, we see that

$$\begin{aligned}
&\int_0^1 |\bar{Q}_{\Gamma_k(\mathbf{s})}(r) - \bar{Q}_{\Gamma_k(\mathbf{s}')}(r)|^p dr \\
&\quad = \int_0^1 |\sum_{i=0}^{M} \bar{\delta}(r,i) \left((a_{\Gamma_k(\mathbf{s}),i} - a_{\Gamma_k(\mathbf{s}'),i})r + b_{\Gamma_k(\mathbf{s}),i} - b_{\Gamma_k(\mathbf{s}'),i} \right)|^p dr \\
&\quad = \sum_{i=0}^{M} \int_{\sigma(\alpha_i)}^{\sigma(\alpha_{i+1})} |\bar{a}_{i,k}r + \bar{b}_{i,k}|^p dr \qquad (11) \\
&\quad = \sum_{i=0}^{M} \frac{(\bar{a}_{i,k}r + \bar{b}_{i,k})|\bar{a}_{i,k}r + \bar{b}_{i,k}|^p}{\bar{a}_{i,k}(p+1)} \Bigg|_{\sigma(\alpha_i)}^{\sigma(\alpha_{i+1})} \qquad (12)
\end{aligned}$$

where in Equation 12 we use the notation $G(r)|_a^b = G(b) - G(a)$. In Equation 11 we integrate over subintervals $[\sigma(\alpha_i), \sigma(\alpha_{i+1})]$ of the interval $[0, 1]$, and can

therefore remove the indicator function $\bar{\delta}(r,i)$. In Equation 12 we solve the integral, where we exploit that according to product and chain rules

$$\begin{aligned}
&\frac{\partial}{\partial r}\frac{(\bar{a}_{i,k}r+\bar{b}_{i,k})|\bar{a}_{i,k}r+\bar{b}_{i,k}|^p}{\bar{a}_{i,k}(p+1)} \\
&\quad= \frac{\bar{a}_{i,k}|\bar{a}_{i,k}r+\bar{b}_{i,k}|^p+(\bar{a}_{i,k}r+\bar{b}_{i,k})p|\bar{a}_{i,k}r+\bar{b}_{i,k}|^{p-1}\text{sign}(\bar{a}_{i,k}r+\bar{b}_{i,k})\bar{a}_{i,k}}{\bar{a}_{i,k}(p+1)} \\
&\quad= |\bar{a}_{i,k}r+\bar{b}_{i,k}|^p.
\end{aligned}$$

The claim directly follows from Equation 12. □

An important note with respect to the distance function $d_p(\Psi_\Gamma(\mathbf{s}),\Psi_\Gamma(\mathbf{s}'))$ is that its closed-form computation given by Proposition 1 allows gradients to be propagated through distance computations (as well as through embedding computations as discussed in Section 3) to the parameters of the model Γ defining the embedding. Moreover, all computations can be expressed using standard building blocks available in common deep learning frameworks, such that all gradients are available through automatic differentiation.

4.2 Loss Function

Deep metric learning trains models with loss functions that drive the model towards minimizing distances between pairs of instances from the same class (positive pairs) while maximizing distances between pairs of instances from different classes (negative pairs). Existing approaches differ in the way negative and positive pairs are selected and the exact formulation of the loss. For example, triplet-based losses as introduced by Schroff et al (2015) compare the distance between an anchor instance and another instance from the same class (positive pair) to the distance between the anchor instance and an instance from a different class (negative pair). However, comparing a positive pair with only a single negative pair does not take into account the distance to other classes and can thereby lead to suboptimal gradients; more recent approaches therefore often consider several negative pairs for each positive pair (Oh Song et al, 2016; Sohn, 2016). Inspired by these approaches, we consider several negative pairs for each positive pair, leading to a loss function of the form

$$\mathcal{L} = \sum_{(\mathbf{s_1},\mathbf{s_2})\in\mathcal{P}} \sum_{\substack{(\mathbf{s}_3,\mathbf{s}_4)\in\mathcal{N} \\ \mathbf{s}_3\in\{\mathbf{s}_1,\mathbf{s}_2\}}} \ell(\mathbf{s}_1,\mathbf{s}_2,\mathbf{s}_3,\mathbf{s}_4)$$

where $\mathcal{P}\subset\mathcal{S}_D\times\mathcal{S}_D$ is a set of positive pairs and $\mathcal{N}\subset\mathcal{S}_D\times\mathcal{S}_D$ is a set of negative pairs of instances, and $\ell(\mathbf{s}_1,\mathbf{s}_2,\mathbf{s}_3,\mathbf{s}_4)$ is a loss function that penalizes cases in which a negative pair $(\mathbf{s}_3,\mathbf{s}_4)$ has smaller distance than a positive pair $(\mathbf{s}_1,\mathbf{s}_2)$. A straightforward linear formulation of the loss would be $\ell(\mathbf{s}_1,\mathbf{s}_2,\mathbf{s}_3,\mathbf{s}_4) = d_p(\Psi_\Gamma(\mathbf{s}_1),\Psi_\Gamma(\mathbf{s}_2)) - d_p(\Psi_\Gamma(\mathbf{s}_3),\Psi_\Gamma(\mathbf{s}_4))$. However, only pairs of pairs that violate the distance criterion should contribute to the loss, leading to $\ell(\mathbf{s}_1,\mathbf{s}_2,\mathbf{s}_3,\mathbf{s}_4) = \max(0, d_p(\Psi_\Gamma(\mathbf{s}_1),\Psi_\Gamma(\mathbf{s}_2)) - d_p(\Psi_\Gamma(\mathbf{s}_3),\Psi_\Gamma(\mathbf{s}_4)))$. We

further replace this loss by a smooth upper bound using log-sum-exp, leading to our final Wasserstein-based loss function

$$\mathcal{L} = \sum_{(\mathbf{s_1},\mathbf{s_2})\in\mathcal{P}} \sum_{\substack{(\mathbf{s}_3,\mathbf{s}_4)\in\mathcal{N} \\ \mathbf{s}_3\in\{\mathbf{s}_1,\mathbf{s}_2\}}} log\left(1 + exp^{d_p(\Psi_\Gamma(\mathbf{s}_1),\Psi_\Gamma(\mathbf{s}_2)) - d_p(\Psi_\Gamma(\mathbf{s}_3),\Psi_\Gamma(\mathbf{s}_4))}\right). \quad (13)$$

Equation 13 is of similar structure as other losses used in the literature, including the angular triplet loss (Wang et al, 2017), the lifted structured loss (Oh Song et al, 2016), and the N-pair loss (Sohn, 2016).

It remains to specify how positive pairs $\mathcal{P}$ and negative pairs $\mathcal{N}$ are sampled for each stochastic gradient descent step. We use the approach of Sohn (2016) for generating $\mathcal{P}$ and $\mathcal{N}$, which has been shown to give state-of-the-art performance in several studies (Sohn, 2016; Wu et al, 2017; Yuan et al, 2017), in particular outperforming triplet-based sampling (Schroff et al, 2015) and lifted structure sampling (Oh Song et al, 2016). The approach constructs a batch of size $2N$ (where N is an adjustable parameter) by sampling from the training data N pairs of instances $\mathcal{P} = \{(\mathbf{s}_i, \mathbf{s}_i^+)\}_{i=1}^N$ from N different classes, such that each pair $(\mathbf{s}_i, \mathbf{s}_i^+)$ is a positive pair from a different class. From the sampled batch, a set of $N(N-1)$ negative pairs is constructed by setting $\mathcal{N} = \{(\mathbf{s}_i, \mathbf{s}_j^+)\}_{\substack{i,j=1 \\ j\neq i}}^N$. Note that Equation 13 can be computed by first computing the embeddings of the $2N$ instances in the batch, and then computing the overall loss. Thus, although the computation is quadratic in N, the number of evaluations of the deep neural network model Γ is linear in the batch size.

5 Empirical Study

We empirically study the proposed method in three biometric identification domains involving human eye movements, accelerometer-based observation of human gait, and EEG recordings. As an ablation study, we specifically evaluate which impact the different components of our proposed method – the metric learning approach, the use of quantile layers to fit the distribution of activations of filters across a sequence, and the Wasserstein-based distance function – have on overall performance.

5.1 Data Sets

We study biometric identification based eye movements, the gait, or the EEG signal of a subject. In all domains, the data consist of sequential observations of the corresponding low-level sensor signal – gaze position from an eye tracker, accelerometer measurements, or EEG measurements – for different subjects. The task is to identify the subject based on the observed sensor measurements.

The *Dynamic Images and Eye Movements* (DIEM) dataset (Mital et al, 2011) contains eye movement data of 210 subjects each viewing a subset of 84 video clips. The video clips are of varying length with an average of 95 seconds

and contain different visual content, such as excerpts from sport matches, documentary videos, movie trailers, or recordings of street scenes. The data contain the gaze position on the screen for the left and the right eye, as well as a measurement of the pupil dilation, at a temporal resolution of 30 Hz. The eye movement data of a particular individual on a particular video clip is thus given by a sequence of six-dimensional vectors (horizontal and vertical gaze coordinate for left and right eye plus left and right pupil dilation), that is, $D = 6$ in the notation of Section 3. The average sequence length is 2850 and there are 5381 sequences overall.

The gait data we use comes from a study by Ihlen et al (2015) who collected the daily movement activity of 71 subjects for a period of 3 consecutive days. The recorded data consists of time series of 3D accelerometer measurements recorded at a sampling rate of 100Hz. For each point in time, the measurement is a $D = 6$ dimensional vector consisting of the acceleration and velocity in x, y, and z direction. In the original data set, a continuous measurement for 3 days has been carried out for each individual. These long measurements contain different activities, but also long idle periods (for example, during sleep). We concentrate on subsequences showing high activity, by dividing the entire recording for each subject into intervals of length one minute, and then selecting for each subject the 30 subsequences that had the largest standard deviation in the 6-dimensional observations. This resulted in 2130 sequences overall (30 for each of the 71 subjects), with a length of $T = 6000$ per sequence.

The EEG data we use come from a study by Zhang et al (1995) who conducted EEG recording sessions with 121 subjects, measuring the signal from 64 electrodes placed on the scalp at a temporal resolution of 256Hz of the subjects while viewing an image stimulus. The original aim of the study was to find a correlation between EEG observations and genetic predisposition to alcoholism, but as subject identifiers are available for all recordings the data can also be used in a biometric setting. Each subject completed between 40 and 120 trials with 1 second of recorded data per trial. The resulting data therefore consist of sequences of $D = 64$ dimensional vectors with a sequence length of 256 (one trial for one subject).

5.2 Problem Setting

As usual in metric learning, we study a setting in which there are distinct sets of subjects at training and test time. The embedding model is first trained on a set of training subjects. On a separate and disjoint set of test subjects, we then evaluate to what degree the learned embedding assigns small distances to pairs of test sequences from the same subject, and large distances to pairs of sequences from different subjects. This reflects an application setting in which new subjects are registered in a database without retraining the embedding model. It also naturally allows the identification of imposters, that is, subjects who have never been observed (neither during training nor in the database of registered subjects) and try to gain access to the system.

In all three domains, we therefore first split the data into training and test data, such that there is no overlap in subjects between the two. For training the embedding model, we use data of 105 of the 210 subjects (eye movements), 36 of 71 subjects (gait data), or 61 of 121 subjects (EEG data). For the eye movement domain, we additionally ensure that there is no overlap in visual stimulus (video clips) between training and test data by splitting the set of all videos into training and test videos and only keeping the respective sequences in the training and test data. During training, each sequence constitutes an instance and the subject its class, and we train either embedding models using metric learning as discussed in Section 4 or, as a baseline, multiclass classification models (see Section 5.3 for details). We also set apart the data of 20% of the training individuals as validation data to tune model hyperparameters.

At test time, we simulate a biometric application setting by first sampling, for each test subject, a random subset of the sequences available for that subject as instances that are put in an enrollment database. We then simulate that we observe additional sequences from a subject which are compared to the sequences of all subjects in the enrollment database. An embedding is good if the distance between these additional sequences and the enrollment sequences of the same subject is low, compared to the distance to the enrollment sequences of other subjects. More precisely, for each subject we use all except five of the sequences available for that subject as enrollment sequences. We then study how well the subject can be identified based on observing n of the remaining sequences, for $n \in \{1, .., 5\}$. Given observed sequences $\mathbf{s}_1, ..., \mathbf{s}_n$ (representing a subject that is unknown at test time), we compute distances to all subjects j as $d_j = \frac{1}{n} \sum_{i=1}^{n} d(\mathbf{s}_i, \mathbf{s}_{ij})$ where $\mathbf{s}_{ij}$ is the sequence of subject j in the enrollment database with minimal distance to $\mathbf{s}_i$. Here, the definition of the distance function d is method-specific (see below for details).

We first study a *verification* scenario. This is the binary problem of deciding if the observed sequences $\mathbf{s}_1, ..., \mathbf{s}_n$ match a particular subject j, by comparing the computed distance d_j to a threshold value. Varying the threshold trades of false-positive versus false-negative classifications, yielding a ROC curve and AUC score. Note that the verification scenario also covers the setting in which in imposter is trying to get access to a system as a particular user; the false-positive rate is the rate at which such imposters would be accepted.

We then study a *multiclass identification* scenario, where we use the model to assign the observed sequences $\mathbf{s}_1, ..., \mathbf{s}_n$ to a subject enrolled in the database (the subject $j^* = \arg\min_j d_j$). This constitutes a multiclass classification problem for which (multiclass) accuracy is measured. In this experiment, we also vary the number of subjects under study, by randomly sampling a subset of subjects which are enrolled in the database; the same subset of subjects is observed at test time. The identification problem becomes more difficult as the number of subjects increases.

We finally study the robustness of the model to imposters in the multiclass identification scenario, an experiment we denote as *multiclass imposters*. This reflects applications in which access to a system does not require a user name, because the system tries to automatically identify who is trying to gain access.

In this experiment, half of the test subjects play the role of imposters who are not registered in the enrollment database. As in the multiclass identification setting, observed sequences are matched to the enrolled subject with minimum distance. This minimum distance is then compared to a threshold value; if the threshold is exceeded, the match is rejected and the observed sequences are classified as belonging to an imposter. Varying the threshold trades off false-positives (match of imposter accepted) versus false-negatives (match of a subject enrolled in the database rejected), yielding a ROC curve and AUC. Correctly rejecting imposters is harder in this setting because it suffices for an imposter to successfully impersonate any enrolled subject. In this experiment we also vary the number of subjects enrolled in the database.

In all three scenarios, the split of sequences into enrollment and observed sequences is repeated 10 times to obtain standard deviations of results. Moreover, accuracies and AUCs will increase with increasing n, as identification becomes easier the more data of an unknown subject is available.

5.3 Methods Under Study

We generally study the deep convolutional architecture proposed by Abdelwahab and Landwehr (2019) for biometric identification, which consists of 16 stacked 1D-convolution layers with PReLU activation functions. We vary the aggregation operation, loss function, and training algorithm in order to evaluate the impact of these components on overall performance.

QP-WL: Our method, combining the quantile embeddings of Section 3 with the Wasserstein-based loss function and metric learning algorithm of Section 4. In all experiments, we set the parameter p of the distance function (see Definition 2) to one, that is, we use the Earth Mover Distance variant of the Wasserstein distance. The convolutional neural network architecture Γ of Section 3 is given by 16 stacked convolution layers with parametric RELU activations as defined by Abdelwahab and Landwehr (2019). The number of sampling points for the quantile function is $M = 16$. At test time, distance between instances is given by the distance function from Definition 2.

QP-NPL: This method uses the same network architecture and quantile embedding as *QP-WL*. However, the resulting quantile embedding is then flattened into an $K \cdot M$ vector embedding, with entries $\bar{Q}_{\Gamma_k(\mathbf{s})}(\sigma(\alpha_m))$ for $k \in \{1, ..., K\}$ and $m \in \{1, ..., M\}$. Then the standard N-pair loss, which is based on cosine similarities between embedding vectors (Sohn, 2016), is used for training. At test time, the distance between instances is given by negative cosine similarity. This method utilizes quantile-based aggregation and metric learning, but does not employ our Wasserstein-based loss function.

MP-NPL: This method uses the same basic network architecture as *QP-NPL*, but uses standard max-pooling instead of a quantile layer for global aggregation. This results in a K-dimensional embedding vector. As for *QP-NPL*, the model is trained using metric learning with the N-pair loss. At test

Eye data	1 Video	2 Videos	3 Videos	4 Videos	5 Videos
QP-WL	**0.9466**±0.0032	**0.9716**±0.0020	**0.9799**±0.0013	**0.9837**±0.0008	**0.9860**±0.0005
QP-NPL	0.9345±0.0033	0.9584±0.0027	0.9667±0.0020	0.9705±0.0014	0.9738±0.0010
MP-NPL	0.8890±0.0035	0.9232±0.0028	0.9334±0.0017	0.9392±0.0014	0.9437±0.0016
QP-CLS	0.9007±0.0053	0.9318±0.0029	0.9424±0.0025	0.9503±0.0025	0.9538±0.0026
Gait data	1 Minute	2 Minutes	3 Minutes	4 Minutes	5 Minutes
QP-WL	**0.9923**±0.0008	**0.9963**±0.0003	**0.9971**±0.0003	**0.9974**±0.0002	**0.9978**±0.0001
QP-NPL	0.9889±0.0009	0.9932±0.0004	0.9943±0.0003	0.9947±0.0002	0.9951±0.0002
MP-NPL	0.9459±0.0027	0.9624±0.0027	0.9690±0.0021	0.9735±0.0016	0.9757±0.0012
QP-CLS	0.9579±0.0040	0.9756±0.0018	0.9812±0.0016	0.9856±0.0011	0.9878±0.0008
EEG data	1 Second	2 Seconds	3 Seconds	4 Seconds	5 Seconds
QP-WL	**0.9968**±0.0006	**0.9985**±0.0001	**0.9988**±0.0001	**0.9991**±0.0000	**0.9992**±0.0000
QP-NPL	0.9927±0.0005	0.9941±0.0005	0.9953±0.0003	0.9955±0.0002	0.9959±0.0001
MP-NPL	0.9611±0.0012	0.9687±0.0005	0.9713±0.0005	0.9722±0.0005	0.9732±0.0005
QP-CLS	0.9796±0.0017	0.9868±0.0009	0.9901±0.0010	0.9920±0.0006	0.9923±0.0007

Table 1 Area under the ROC curve with standard error for all methods and domains in the verification setting for varying number $n \in \{1, 2, 3, 4, 5\}$ of observed sequences.

time, distance is given by negative cosine similarity. This baseline uses metric learning, but neither quantile layers nor the Wasserstein-based loss function.
QP-CLS: This baseline uses the same network architecture and flattened quantile embedding as *QP-NPL*, but feeds the flattened embedding vector into a dense classification layer with softmax activation. The models is trained in a classification setting using multiclass crossentropy. Distance at test time is given by negative cosine similarity. This model is identical to the model presented in Abdelwahab and Landwehr (2019), except that we remove the final classification layer at test time to generate embeddings for novel subjects.

For all methods, training is carried out using the Adam optimizer with learning rate 0.0001 for 50000 iterations, and the regularizer of the PReLU activation function is tuned as a hyperparameter on the validation set as in (Abdelwahab and Landwehr, 2019).

5.4 Results

We present and discuss empirical results for the different domains in turn.

5.4.1 Eye Movements

Table 1, upper third, shows area under the ROC curve for all methods and varying number n of observed sequences in the eye movement domain. Comparing *QP-WL* and *QP-NPL*, we observe that the Wasserstein-based loss introduced in Section 4, which works on the distributional embedding given by the piecewise linear approximations of the quantile functions, clearly outperforms flattening the distributional embedding and using N-pair loss. Comparing *MP-NPL* with *QP-NPL* and *QP-WL* shows that using quantile layers

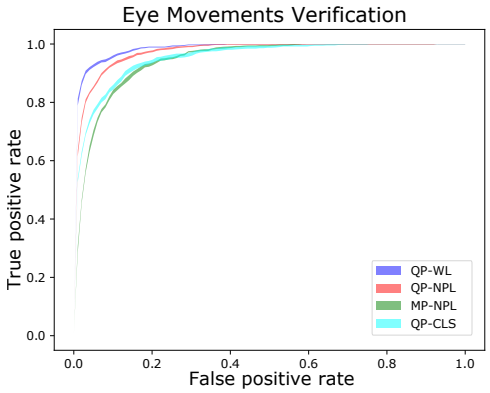

Fig. 2 ROC curves in the eye movement domain for all methods using $n = 5$ observed sequences. Shaded region in ROC curves indicates standard error.

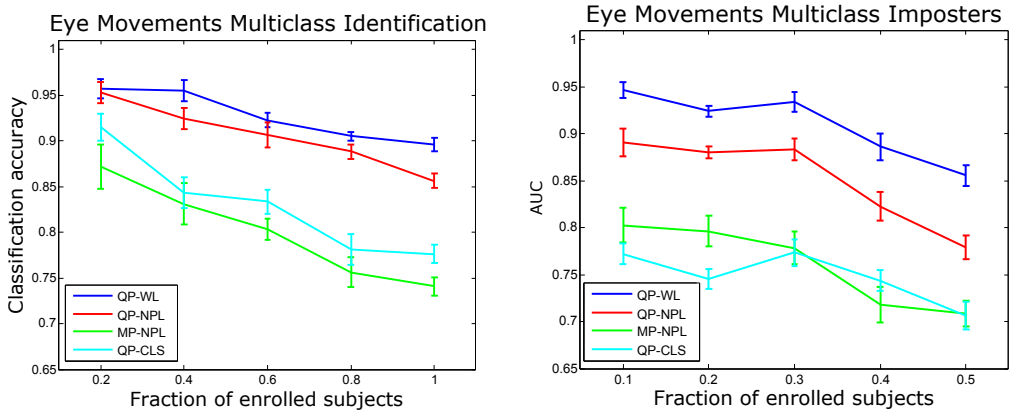

Fig. 3 Left: Identification accuracy in the multiclass identification scenario for the eye movement domain and $n = 5$ observed test instances as a function of the fraction of subjects that are enrolled. Right: area under the ROC curve for multiclass imposters as a function of the fraction of subjects enrolled. In the imposter scenario, 50% of subjects are imposters and therefore never enrolled. Error bars indicate the standard error.

improves accuracy compared to max-pooling even if the quantile embedding is flattened (and more so if Wasserstein-based loss is used). Classification training (*QP-CLS*) reduces accuracy compared to metric learning (*QP-NPL*). As expected, AUC increases with the number n of sequences observed at test time. Figure 2 shows ROC curves in the verification setting for $n = 5$.

Figure 3 (left) shows multiclass identification accuracy for $n = 5$ observed sequences as a function of the fraction of the 105 subjects who are enrolled. Relative results for the different methods are similar as in the verification setting. Accuracy decreases slightly when more subjects are enrolled, as the multiclass problem becomes more difficult. Figure 3 (right) shows the robustness of the model to multiclass imposters as a function of the fraction of the 105 subjects who are enrolled (up to 50%, as half of the subjects are imposters). We observe that *QP-WL* is much more robust to imposters than the baseline methods.

In the eye movement domain, we also compare against the state-of-the-art model by Jager et al (2019), denoted *Jager et al. (2019)*. *Jager et al. (2019)* uses angular gaze velocities averaged over left and right eye as input, which we

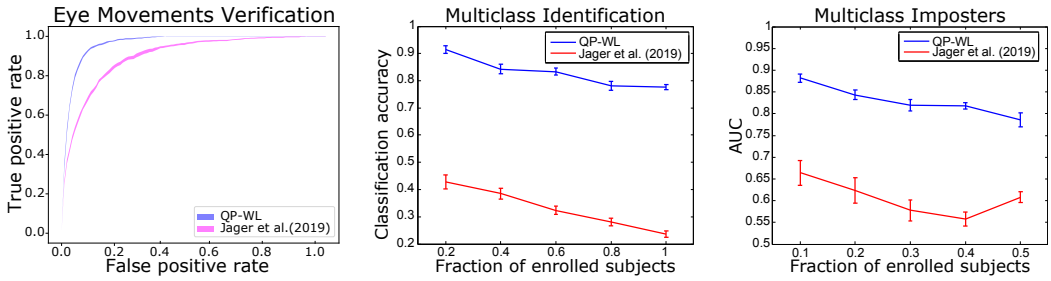

Fig. 4 Comparison between *QP-WL* and *Jager et al. (2019)* in the eye movement domain: area under ROC curve in verification scenario (left), identification accuracy in multiclass identification scenario (center), and robustness of model to multiclass imposters (right). In this experiment, the data is simplified for both methods to match the requirements of *Jager et al. (2019)*, see text for details. Results of *QP-WL* therefore differ from results presented in Figure 2 and Figure 3. Error bars indicate the standard error.

compute from our raw data. We replicate the setting of Jager et al (2019) by training the model using multiclass classification and using the last layer before the classification layer as the embedding at test time. The *Jager et al. (2019)* architecture cannot deal with variable-lenght sequences, we therefore split the variable-length sequences in our data into shorter sequences of fixed length, namely the length of the shortest sequence (27 seconds). For a fair comparison, we also simplify the data for our model in this experiment: using only the average gaze point rather than left and right gaze point separately, removing pupil dilation, and using the same fixed-length sequences. Figure 4 shows ROC curves for the verification scenario (left) and identification accuracy (center) as well as AUC in the imposter scenario for our model *QP-WL* on the simplified data and *Jager et al. (2019)*. Comparing to Figure 2 and Figure 3 we observe that accuracies are reduced for our model by using the simplified data, but the model still outperforms *Jager et al. (2019)* by a wide margin. We note that the model of Jager et al (2019) is focused on microsaccades, which are likely not detectable in our data due to the low temporal resolution (30Hz compared to 1000Hz in the study by Jager et al (2019)), which might explain the relatively poor performance of the model on our data.

5.4.2 Gait

Table 1, center third, shows area under the ROC curve for all methods and varying number n of observed sequences in the gait domain. We observe the ordering in terms of relative performance between the different methods as in the eye movements domain, with clear benefits when using the proposed loss function based on Wasserstein distance (*QP-WL* versus *QP-NPL*), when using quantile layers instead of max-pooling aggregation (*QP-WL* and *QP-NPL* versus *MP-NPL*), and when using metric learning rather than classification-based training (*QP-NPL* versus *QP-CLS*). Figure 5 shows ROC curves for verification at $n = 5$ in the gait domain. Figure 6 (left) shows identification accuracy as a function of the fraction of subjects enrolled in the gait domain; in this setup the ordering of methods in terms of performance is the same

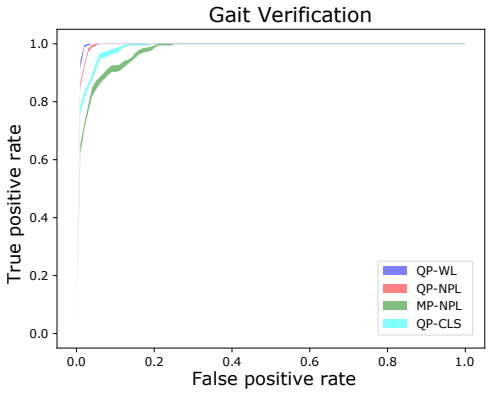

Fig. 5 ROC curves in the gait domain for all methods using $n = 5$ observed sequences. Shaded region in ROC curves indicates standard error.

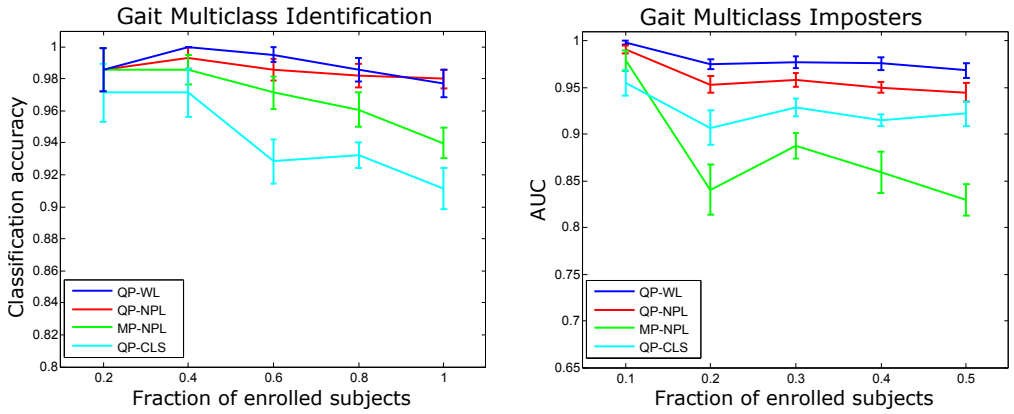

Fig. 6 Left: Identification accuracy in the multiclass identification scenario for the gait domain and $n = 5$ observed test instances as a function of the fraction of subjects that are enrolled. Right: area under the ROC curve for multiclass imposters as a function of the fraction of subjects enrolled. In the imposter scenario, 50% of subjects are imposters and therefore never enrolled. Error bars indicate the standard error.

but the difference between *QP-WL* and *QP-NPL* less pronounced. Figure 6 (right) shows robustness to multiclass imposters, with again a clear advantage of *QP-WL* over the baselines.

5.4.3 EEG

Table 1, bottom third, shows area under the ROC curve for all methods and varying number n of observed test sequences in the EEG domain. Relative performance of methods is generally similar as in the other two domains. *QP-WL* clearly outperforms the closest baseline, reducing 1-AUC by between 56% ($n = 1$) and 80% ($n = 5$). Figure 7 shows ROC curves in the verification setting. Figure 8 (left) and Figure 8 (right) show identification accuracy as a function of the fraction of subjects enrolled and robustness of the models to multiclass imposters. As in the gait domain, differences are more pronounced in the latter setting.

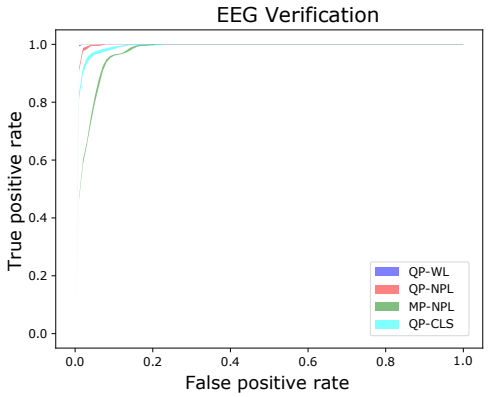

Fig. 7 ROC curves in the EEG domain for all methods using $n = 5$ observed sequences. Shaded region in ROC curves indicates standard error.

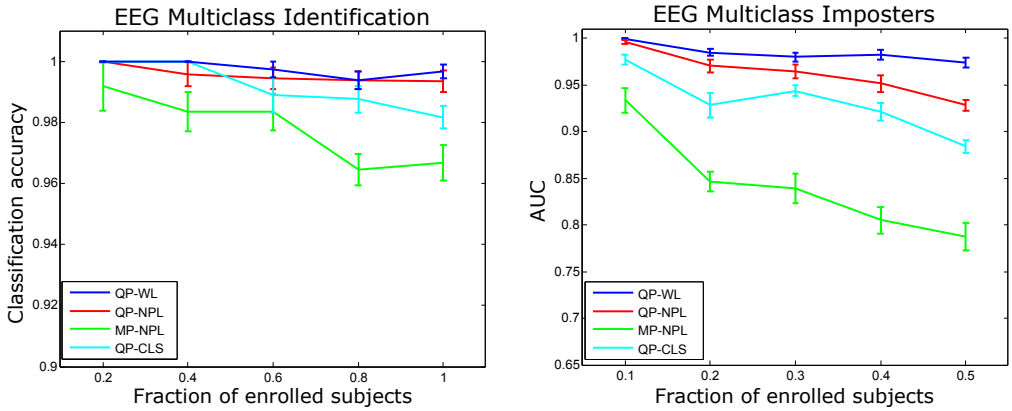

Fig. 8 Left: Identification accuracy in the multiclass identification scenario for the EEG domain and $n = 5$ observed test instances as a function of the fraction of subjects that are enrolled. Right: area under the ROC curve for multiclass imposters as a function of the fraction of subjects enrolled. In the imposter scenario, 50% of subjects are imposters and therefore never enrolled. Error bars indicate the standard error.

6 Conclusions

We developed a model for distributional embeddings of variable-length sequences using deep neural networks. Building on existing work on quantile layers, the model represents an instance by the distribution of the learned deep features across the sequence. We developed a distance function for these distributional embeddings based on the Wasserstein distance between the corresponding distributions, and from this distance function a loss function for performing metric learning with the proposed model. A key point about the model is end-to-end learnability: by using piecewise linear approximations of the quantile functions, and based on those providing a closed-form solution for the Wasserstein distance, gradients can be traced through the embedding and loss calculations. In our empirical study, distributional embeddings outperformed standard vector embeddings by a large margin on three data sets from different domains.

References

Abdelwahab A, Landwehr N (2019) Quantile layers: Statistical aggregation in deep neural networks for eye movement biometrics. In: Proceedings of the 30th European Conference on Machine Learning

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: International conference on machine learning, pp 214–223

Athiwaratkun B, Wilson A (2017) Multimodal word distributions. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 1645–1656

Bojchevski A, Günnemann S (2018) Deep Gaussian embedding of graphs: Unsupervised inductive learning via ranking. In: International Conference on Learning Representations, pp 1–13

Bucher M, Herbin S, Jurie F (2016) Improving semantic embedding consistency by metric learning for zero-shot classiffication. In: European Conference on Computer Vision, Springer, pp 730–746

Cambanis S, Simons G, Stout W (1976) Inequalities for ek (x, y) when the marginals are fixed. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 36(4):285–294

Chung JS, Nagrani A, Zisserman A (2018) Voxceleb2: Deep speaker recognition. Proc Interspeech 2018 pp 1086–1090

Frogner C, Zhang C, Mobahi H, Araya M, Poggio TA (2015) Learning with a Wasserstein loss. In: Advances in Neural Information Processing Systems, pp 2053–2061

Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with Wasserstein distance. arXiv preprint arXiv:160402199

Gibiansky A, Arik S, Diamos G, Miller J, Peng K, Ping W, Raiman J, Zhou Y (2017) Deep voice 2: Multi-speaker neural text-to-speech. In: Advances in neural information processing systems, pp 2962–2970

Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, vol 2, pp 1735–1742

Ihlen EA, Weiss A, Helbostad JL, Hausdorff JM (2015) The discriminant value of phase-dependent local dynamic stability of daily life walking in older adult community-dwelling fallers and nonfallers. BioMed research international

Jager L, Makowski S, Prasse P, Liehr S, Seidler M, Scheffer T (2019) Deep eyedentification: Biometric identification using micro-movements of the eye. In: Proceedings of the 30th European Conference on Machine Learning

Li C, Ma X, Jiang B, Li X, Zhang X, Liu X, Cao Y, Kannan A, Zhu Z (2017) Deep speaker: an end-to-end neural speaker embedding system. arXiv preprint arXiv:170502304

McLaughlin N, Martinez del Rincon J, Miller P (2016) Recurrent convolutional network for video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1325–1334

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances

in neural information processing systems, pp 3111–3119
Mital PK, Smith TJ, Hill RL, Henderson JM (2011) Clustering of gaze during dynamic scene viewing is predicted by motion. Cognitive Computation 3(1):5–24
Mueller J, Thyagarajan A (2016) Siamese recurrent architectures for learning sentence similarity. In: Thirtieth AAAI Conference on Artificial Intelligence
Neculoiu P, Versteegh M, Rotaru M (2016) Learning text similarity with siamese recurrent networks. In: Proceedings of the 1st Workshop on Representation Learning for NLP, pp 148–157
Oh Song H, Xiang Y, Jegelka S, Savarese S (2016) Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4004–4012
Reed S, Akata Z, Lee H, Schiele B (2016) Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 49–58
Resnick SI (2013) Extreme values, regular variation and point processes. Springer
Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
Sedighi V, Fridrich J (2017) Histogram layer, moving convolutional neural networks towards feature-based steganalysis. Electronic Imaging 2017(7):50–55
Sohn K (2016) Improved deep metric learning with multi-class n-pair loss objective. In: Advances in Neural Information Processing Systems, pp 1857–1865
Vilnis L, McCallum A (2015) Word representations via Gaussian embedding. International Conference on Learning Representations (ICLR)
Wang J, Zhou F, Wen S, Liu X, Lin Y (2017) Deep metric learning with angular loss. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2593–2601
Wang Z, Li H, Ouyang W, Wang X (2016) Learnable histogram: Statistical context features for deep neural networks. In: European Conference on Computer Vision, Springer, pp 246–262
Wu CY, Manmatha R, Smola AJ, Krahenbuhl P (2017) Sampling matters in deep embedding learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2840–2848
Wu L, Wang Y, Gao J, Li X (2018) Where-and-when to look: Deep siamese attention networks for video-based person re-identification. IEEE Transactions on Multimedia 21(6):1412–1424
Yuan Y, Yang K, Zhang C (2017) Hard-aware deeply cascaded embedding. In: Proceedings of the IEEE international conference on computer vision, pp 814–823
Zhang XL, Begleiter H, Porjesz B, Wang W, Litke A (1995) Event related potentials during object recognition tasks. Brain Research Bulletin 38(6):531–538

Chapter 5

Discussion

This thesis has studied machine learning approaches for eye movement biometrics. On the one hand, it has shown that the proposed novel machine learning models can improve the identification accuracy of eye movement biometrics in various settings. On the other hand, the developed machine learning approaches — especially those of Chapters 3 and 4 — are generally applicable, and thus also generally contribute to the machine learning literature, especially the deep learning and deep metric learning literature. Empirically, this thesis has systematically studied the performance of the proposed models and baselines on several large-scale eye movement data sets, comprising 461 individuals and 38,993 gaze sequences overall.

At first, the thesis introduced a Bayesian semiparametric model for eye movement biometrics. On the application side, this model improves over earlier fully parametric models for eye movement biometrics. On the methodological side, the study has introduced a solution for performing Bayesian inference with models involving observation-specific truncations of densities. Next, the thesis introduced a statistical aggregation layer to fit distributions of learned deep features within neural networks. Empirically, using the proposed deep neural networks instead of probabilistic models built upon psychological concepts improves biometric identification accuracy. Moreover, deep neural networks with the proposed statistical aggregation layer outperform networks with standard pooling layers and the other statistical aggregation layers proposed in the literature. Finally, in Chapter 4, the thesis introduced a novel deep metric learning approach for sequence data, in which input sequences are represented by the distributions of learned deep features across the sequence. This approach is again general and not limited to eye movements, having been empirically shown to outperform existing deep metric learning techniques in several domains. Before discussing the biometrical empirical results in Section 5.2, I first discuss the related work in Section 5.1. The contributions of this thesis to the field of machine learning and eye movement modeling are discussed in Section 5.3.

5.1 Related Work

This thesis starts by proposing the semiparametric approach for modeling eye movement sequences using a model based on the psychological concepts of eye movement. The model is specifically based on the concepts for eye movements during reading, distinguishing different saccade types [Heister 12]. In an effort to understand eye movements during reading, several studies analyzed gaze paths based on the text

content and structure, using machine learning techniques to predict the next gaze movement [Hara 12, Matthies 13]. When developing Chapter 2, a number of studies addressed the problem of reader identification from eye movement while reading arbitrary text [Landwehr 14, Holland 11]. One study used handcrafted aggregated features across the gaze movement sequence to describe the sequence [Holland 11]. After computing similarities between these features, the overall similarity between two given sequences was established by summing these similarities. In that study, the contribution of each feature to the overall similarity was weighted. The majority of the weights (more than two-thirds) were given to only two features: average fixation duration and the number of fixations. In comparison to this feature-based model, better performance was demonstrated for a probabilistic model that captures individual differences in distributions over the fixation durations and saccade amplitudes of the gaze sequences generated by an individual [Landwehr 14]. Specifically, in this model, saccadic moves were segmented into four groups: *refixation*, *next word movement*, *forward skip*, and *regression*. The distributions of fixation durations and saccade amplitudes over each movement group were modeled separately for each individual. As some of these distributions were estimated from sparse data, and to avoid overfitting, a simple parametric model based on Gamma distributions was used. This thesis compares the semiparametric model proposed in Chapter 2 with different variations of both models [Holland 11, Landwehr 14]. There is also general work on semiparametric density estimation in statistics and machine learning [Yang 09, Lenk 03, Hjort 95]. However, these approaches are not directly applicable to the distributions in our domain as they are truncated at each observation differently depending on the structure of the input text.

Psychological concepts have so far been the foundation for most eye movement biometric approaches, as well as for modeling eye movements on non-textual visual input. Previous studies, not limited to textual input, focused on segmenting the gaze movement sequences into fixations and saccades. Using fixations and saccades, the studies mainly designed handcrafted features that could differentiate between sequences generated by different individuals. Simple features such as average fixation durations and average saccade amplitudes were used for an identification study [Holland 13]. This work was later extended to include more complex features like saccade acceleration, using statistical tests to compare sequences [Rigas 16]. Another study used probabilistic modeling [Kinnunen 10], wherein the authors slid a window over the gaze movement sequence, after preprocessing the gaze coordinates into angles between consecutive gaze points. The authors computed a histogram describing the angles in each window. Assuming independence between the windows, the authors built an individual-specific distribution as a mixture of multivariate Gaussians. Jointly, existing work in the literature has shown that statistically modeling short-term local patterns (fixations and saccades) best characterizes an individual. Instead of preprocessing the data and using handcrafted features, a deep neural network is introduced in Chapter 3. The network extracts the short-term local patterns with stacked convolution layers, then characterizes their distribution using a statistical aggregation layer. This chapter compares the proposed method with all the related methods [Holland 13, Rigas 16, Kinnunen 10, Landwehr 14] and the prior work in Chapter 2, illustrating that the proposed method outperforms them by a wide margin.

Other statistical aggregation layers for deep neural networks have been proposed

in the literature [Wang 16, Sedighi 17]. The layers are based on approximating the histogram of learned deep features within the network. The layers also contain trainable parameters, namely the bin centers and widths within the histogram. One study used a linear approximation of the histogram to enable the gradient to flow [Wang 16], while the other used a Gaussian kernel as an approximation for the histogram bin [Sedighi 17]. Unlike the quantile layer introduced in Chapter 3, they do not generalize standard pooling layers. Empirically, this thesis compares these existing statistical aggregation layers to the quantile layer and shows that the proposed quantile layer outperforms these histogram-based layers. In the empirical study presented in Chapter 3, the proposed model is compared with a recurrent architecture based on gated recurrent units [Cho 14], which are a popular class of recurrent architectures, as they have shown to be more robust and faster to train than LSTMs while giving similar performance [Chung 14].

The deep metric learning approach studied in Chapter 4 is related to other work in deep metric learning for sequences. A number of studies have worked with sequence deep metric learning. An earlier study [Chung 18] proposed an embedding for voice sequences in two stages. First, the network was trained for classification using a softmax loss, then it was fine-tuned with contrastive loss [Hadsell 06], using Euclidean distance to compare embeddings. An alternative study [Li 17] used triplet loss [Schroff 15] to train a speaker identification model, using cosine similarity to compare embeddings. Another work [Mueller 16] introduced LSTM-based siamese networks [Bromley 94] to calculate the similarity between two sentences, using the Manhattan distance to compare embeddings. In Chapter 4, the thesis studied the embeddings of sequences based on distributions. The existing deep metric learning distance functions proposed in the literature are not directly applicable to measuring the distances between distributional embeddings. There is also a rich literature on loss functions for deep metric learning. Several studies [Sohn 16, Wu 17, Yuan 17] illustrate that N-pair loss [Sohn 16] is a state-of-the-art loss function for deep metric learning, outperforming lifted structure [Oh Song 16] and triplet loss [Schroff 15]. Chapter 4 compares the proposed Wasserstein based loss with the N-pair loss.

All the previously mentioned metric learning methods have studied a similar supervised setting as the one studied in the thesis but employed (non-distributional) fixed-point vector embeddings. Other studies have explored the distributional embeddings but in an unsupervised way. Distributional word embeddings were studied in the field of natural language processing, representing an instance as a Gaussian distribution [Vilnis 15] or mixtures of Gaussian distributions [Athiwaratkun 17]. A similar approach proposed distributional embeddings for graph nodes based on Gaussian distributions [Bojchevski 18]. The objective of these approaches was to find a low dimensional embedding space that preserves the distribution of the neighbors for each instance, as in the original space. The distribution in these models does not capture the inner structure of the instance being mapped (distribution of deep features across the sequence), but rather captures uncertainty about the embedding. A further difference with the thesis proposed method is that the proposed method does not assume any fixed parametric form (such as Gaussian) for the distribution, but directly approximates its quantile function.

Concurrently with the work developed in the thesis, a new study [Jäger 19] explored the use of deep neural networks for eye movement biometrics. It proposed a deep neural network framework for reader identification from micro eye movements

after preprocessing the gaze coordinates into gaze angular velocity. The study then divided the angular speed into fast and slow velocity to be processed by two different neural networks. After training each network separately for identification, the authors combined both networks, then trained the concatenating layer. The new study used the intermediate layer as an embedding value for the input. The new study is empirically compared with the approach proposed in Chapter 4.

5.2 Empirical Evaluation

This section summarizes the empirical performance in eye movement biometrics obtained in this thesis, situating it with respect to the literature. Chapter 2 of the thesis introduced a novel probabilistic model for the eye movements observed during reading. The method was empirically evaluated and compared to baseline methods from the literature on a dataset of 251 readers, wherein each read between 100–144 sentences from the Potsdam sentence corpus [Kliegl 06]. For arbitrary test inputs different from the training inputs, the corpus sentences were split randomly into equal sets: training and test sentences. Two settings were studied: multi-class classification setting (identification) and binary classification setting (verification). In the multi-class classification setting, the proposed method in Chapter 2 reduced the error by more than a factor of three compared to the state of the art in reader identification [Landwehr 14]. An even better performance gain is observed compared to the feature-based model [Holland 11]. For the multi-class classification problem, the effect of the number of readers on the performance of the different methods was measured. Another parameter that contributes to the identification error is the number of test sentences read by the reader before identification. As the number of sentences read at test time increases, the models become more certain about the identity of the reader; however, a reader thus needs to be observed for longer before his or her identity can be inferred. In all these experiments, the proposed model substantially outperforms the baselines in terms of identification accuracy. In the binary classification or verification setting, the results were presented as a plot between false reject and false accept rates. The study summarized the results in a table for the area under the curve of the different methods. The study evaluated the different methods in the binary classification setting while varying the number of test sentences read by the unknown reader. The proposed method also kept its lead in this setting by a significant margin. In another study [Makowski 18], the proposed semiparametric approach was compared with the parametric approach [Landwehr 14], and this lead was demonstrated on yet another dataset of 61 participants reading a full document at a time.

The thesis then transitioned from studying eye movements during reading to studying eye movements with more general visual inputs. In Chapter 3, the proposed method was evaluated on a large-scale dataset called DIEM [Mital 11]. The dataset contains gaze movement recordings of 210 participants, viewing between 25–26 short clips from a list of 84 videos of different visual content and varying length. The gaze movement time-series were sampled at 30 Hz, which is an extremely low temporal resolution compared to other studies in eye movement biometrics. Chapter 3 introduced the quantile layer — a statistical aggregation layer — to fit the distributions of learned patterns within deep neural networks. The study used the same deep neural network

convolution architecture for all the deep learning approaches but with different global pooling methods: the standard pooling layer (Global max-pooling, Global average pooling), proposed quantile layer, and other statistical pooling layers from the literature [Wang 16, Sedighi 17]. In addition to the convolution architecture, the chapter studied recurrent architecture using Gated Recurrent Units [Cho 14]. To test on arbitrary input, the test inputs should be different from training inputs. The set of videos was split evenly between training and test videos, and the models were trained to differentiate between the gaze movements of different subjects. The multi-class classification identification errors were recorded while varying the number of videos seen by the unknown subjects. As the number of videos watched increases, the problem gets easier, and the models become more certain about the identity of the unknown subject. The results illustrate that the proposed quantile layer consistently outperforms all the other methods. Furthermore, the study compared the new approach with the previous model in Chapter 2 [Abdelwahab 16] and several models proposed in the eye movement biometrics literature [Kinnunen 10, Rigas 12, Holland 13, Landwehr 14]. The proposed method outperformed the state-of-the-art methods in terms of identification accuracy by a wide margin.

Later, this thesis widened the domain of experiments to include all sequences, not just eye movement sequences. In Chapter 4, the study used three different datasets from different domains to validate the proposed approach. In addition to the gaze movement DIEM dataset [Mital 11], the approach was evaluated on a collection of 3D accelerometer recordings of 71 individuals [Ihlen 15], and another collection of electroencephalography(EEG) recordings of 121 individuals [Zhang 95]. Each dataset contained a list of time-series sequences generated by individuals. The study was interested in learning an embedding of these sequences into a metric space that could be used to identify or verify subjects from the observed sequences, even if the subjects did not appear in the training data. For all of the datasets, two settings were studied: the multi-class classification and verification settings. The chapter measured the performance of the multi-class classification with two measures: the identification error and the AUC of identifying an imposter as an enrolled individual versus identifying an enrolled individual as enrolled. These two measures were calculated while varying the number of enrolled individuals. In the verification case, the results are shown as an ROC curve, where true positives are subjects correctly matched and false positives are cases where a subject is matched against a different subject by the model. The chapter compares the proposed loss function with the N-pair loss [Sohn 16] using two different embeddings, with either global maximum pooling features or quantile features as the embedding. The results illustrate that the proposed loss function consistently outperforms the state-of-the-art loss (N-pair loss) for the different embeddings. Furthermore, the results emphasize the gains of using the quantile pooling layer over a standard pooling layer across different domains. Moreover, the study demonstrates the power of metric learning by showing that using the quantile layer output with metric learning outperforms its output without metric learning by a wide margin. In the field of eye movement biometrics, the proposed method has a large performance gain over the latest state of the art in the eye movement biometrics field [Jäger 19] across all experimental settings.

5.3 Contributions to Machine Learning and Eye Movements Modeling

The thesis has studied novel models for eye movement biometrics and novel machine learning approaches motivated by this application. This section summarizes the main contributions in terms of novel models and machine learning methods.

Chapter 2 introduced a novel Bayesian semiparametric model of human eye movements during reading based on Gaussian process density estimation. The main methodological challenge in the model is to perform inference with densities that are truncated differently for each observation, which are caused by the structure of the text being read. The thesis developed an inference algorithm based on Metropolis-Hastings sampling while using 2-point Newton-Cotes quadrature to normalize the truncated likelihoods for each observation. In terms of eye movement biometrics, the proposed method extends the state-of-the-art probabilistic model [Landwehr 14] with a semiparametric approach, which can capture individual differences more accurately than the more rigid fully parametric formulation. Empirically, this method reduces identification errors for eye movement biometrics based on reading by more than a factor of three in a large-scale study of 251 readers and 33,612 sequences.

Chapter 3 introduced a novel statistical aggregation layer for deep neural networks that approximates the distribution of learned deep features across a sequence. Specifically, the layer approximates the quantile function of the distribution, sampled at a set of learnable parameters. The proposed layer generalizes standard pooling layers but provides a more expressive characterization of the distribution of feature activations than a simple maximum or average. Empirically, the proposed quantile layer outperforms other statistical aggregation layers based on histograms that have been proposed in the literature. Moreover, these existing layers do not generalize standard pooling layers. The proposed method has been empirically studied on a large-scale data set of eye movements of subjects viewing arbitrary non-textual inputs, comprising a total of 151 hours of eye movement data. The method demonstrates a large gain in the performance over the state-of-the-art methods in eye movement biometrics, including the previous work introduced in Chapter 2. Furthermore, Chapter 3 [Abdelwahab 19b] provided the first study in the field of eye movement biometrics to apply deep learning directly to the raw data without preprocessing, in an end-to-end learnable approach. Concurrently, a deep learning model for identification from micro eye movements [Jäger 19] was developed. The thesis compared this model with the metric learning method introduced in Chapter 4.

Chapter 4 introduced a distributional deep metric learning approach for sequences wherein the embedding of a sequence is given by a set of distributions. While distributional embeddings have been studied before in the literature [Vilnis 15, Bojchevski 18, Athiwaratkun 17], the method proposed in this chapter is the first such approach for supervised metric learning. Moreover, the learned distributions are not limited to a fixed parametric form. Models based on the proposed loss function outperform models trained with the N-pair loss — a state-of-the-art loss function in deep metric learning [Sohn 16] — in three different datasets from different domains. For the domain of eye movement biometrics, the proposed method demonstrates a large gain in performance compared to the most recent study in the field [Jäger 19] and to our previous model presented in Chapter 3.

Bibliography

[Abdelwahab 16] Ahmed Abdelwahab, Reinhold Kliegl & Niels Landwehr. *A Semi-parametric Model for Bayesian Reader Identification*. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 585–594, 2016.

[Abdelwahab 19a] Ahmed Abdelwahab & Niels Landwehr. *Deep Distributional Sequence Embeddings Based on a Wasserstein Loss*. 2019.

[Abdelwahab 19b] Ahmed Abdelwahab & Niels Landwehr. *Quantile Layers: Statistical Aggregation in Deep Neural Networks for Eye Movement Biometrics*. In Proceedings of the 30th European Conference on Machine Learning. Springer, 2019.

[Afflerbach 15] Peter Afflerbach. Handbook of individual differences in reading: Reader, text, and context. Routledge, 2015.

[Athiwaratkun 17] Ben Athiwaratkun & Andrew Wilson. *Multimodal Word Distributions*. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1645–1656, 2017.

[Bojchevski 18] Aleksandar Bojchevski & Stephan Günnemann. *Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking*. In International Conference on Learning Representations, pages 1–13, 2018.

[Bromley 94] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger & Roopak Shah. *Signature verification using a" siamese" time delay neural network*. In Advances in neural information processing systems, pages 737–744, 1994.

[Cho 14] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk & Yoshua Bengio. *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, 2014.

[Chung 14] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho & Yoshua Bengio. *Empirical evaluation of gated recurrent neural networks on sequence modeling*. In NIPS 2014 Workshop on Deep Learning, December 2014, 2014.

[Chung 18] Joon Son Chung, Arsha Nagrani & Andrew Zisserman. *VoxCeleb2: Deep Speaker Recognition.* Proc. Interspeech 2018, pages 1086–1090, 2018.

[Dixon 51] W. Robert Dixon. *Studies in the psychology of reading.* In W. S. Morse, P. A. Ballantine & W. R. Dixon, editeurs, Univ. of Michigan Monographs in Education No. 4. Univ. of Michigan Press, 1951.

[Duchowski 02] Andrew T Duchowski. *A breadth-first survey of eye-tracking applications.* Behavior Research Methods, Instruments, & Computers, vol. 34, no. 4, pages 455–470, 2002.

[Hadsell 06] Raia Hadsell, Sumit Chopra & Yann LeCun. *Dimensionality reduction by learning an invariant mapping.* In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 1735–1742. IEEE, 2006.

[Hara 12] Tadayoshi Hara, Daichi Mochihashi, Yoshinobu Kano & Akiko Aizawa. *Predicting word fixations in text with a CRF model for capturing general reading strategies among readers.* In Proceedings of the First Workshop on Eye-Tracking and Natural Language Processing, pages 55–70, 2012.

[Heister 12] Julian Heister, Kay-Michael Würzner & Reinhold Kliegl. *Analysing large datasets of eye movements during reading.* Visual word recognition, vol. 2, pages 102–130, 2012.

[Henderson 03] John M Henderson. *Human gaze control during real-world scene perception.* Trends in cognitive sciences, vol. 7, no. 11, pages 498–504, 2003.

[Hjort 95] Nils Lid Hjort & Ingrid K Glad. *Nonparametric density estimation with a parametric start.* The Annals of Statistics, pages 882–904, 1995.

[Holland 11] Corey Holland & Oleg V Komogortsev. *Biometric identification via eye movement scanpaths in reading.* In 2011 International joint conference on biometrics (IJCB), pages 1–8. IEEE, 2011.

[Holland 13] Corey D Holland & Oleg V Komogortsev. *Complex eye movement pattern biometrics: Analyzing fixations and saccades.* In 2013 International conference on biometrics (ICB), pages 1–8. IEEE, 2013.

[Huey 08] Edmund B. Huey. The psychology and pedagogy of reading. Cambridge, Mass.: MIT Press, 1908.

[Ihlen 15] Espen AF Ihlen, Aner Weiss, Jorunn L Helbostad & Jeffrey M Hausdorff. *The discriminant value of phase-dependent local dynamic stability of daily life walking in older adult community-dwelling fallers and nonfallers.* BioMed research international, vol. 2015, 2015.

[Jäger 19] Lena A Jäger, Silvia Makowski, Paul Prasse, Sascha Liehr, Maximilian Seidler & Tobias Scheffer. *Deep Eyedentification: Biometric Identification using Micro-Movements of the Eye.* In Proceedings of the 30th European Conference on Machine Learning. Springer, 2019.

[Kinnunen 10] Tomi Kinnunen, Filip Sedlak & Roman Bednarik. *Towards task-independent person authentication using eye movement signals.* In Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, pages 187–190. ACM, 2010.

[Kliegl 06] Reinhold Kliegl, Antje Nuthmann & Ralf Engbert. *Tracking the mind during reading: The influence of past, present, and future words on fixation durations.* Journal of experimental psychology: General, vol. 135, no. 1, page 12, 2006.

[Landwehr 14] Niels Landwehr, Sebastian Arzt, Tobias Scheffer & Reinhold Kliegl. *A model of individual differences in gaze control during reading.* In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1810–1815, 2014.

[Lenk 03] Peter J Lenk. *Bayesian semiparametric density estimation and model verification using a logistic–Gaussian process.* Journal of Computational and Graphical Statistics, vol. 12, no. 3, pages 548–565, 2003.

[Li 17] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan & Zhenyao Zhu. *Deep speaker: an end-to-end neural speaker embedding system.* arXiv preprint arXiv:1705.02304, 2017.

[Liversedge 00] Simon P Liversedge & John M Findlay. *Saccadic eye movements and cognition.* Trends in cognitive sciences, vol. 4, no. 1, pages 6–14, 2000.

[Makowski 18] Silvia Makowski, Lena A Jäger, Ahmed Abdelwahab, Niels Landwehr & Tobias Scheffer. *A Discriminative Model for Identifying Readers and Assessing Text Comprehension from Eye Movements.* In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 209–225. Springer, 2018.

[Matthies 13] Franz Matthies & Anders Søgaard. *With blinkers on: Robust prediction of eye movements across readers.* In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 803–807, 2013.

[Mital 11] Parag K Mital, Tim J Smith, Robin L Hill & John M Henderson. *Clustering of gaze during dynamic scene viewing is predicted by motion.* Cognitive Computation, vol. 3, no. 1, pages 5–24, 2011.

[Mueller 16] Jonas Mueller & Aditya Thyagarajan. *Siamese recurrent architectures for learning sentence similarity.* In Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[Oh Song 16] Hyun Oh Song, Yu Xiang, Stefanie Jegelka & Silvio Savarese. *Deep metric learning via lifted structured feature embedding.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4004–4012, 2016.

[Poynter 13] William Poynter, Megan Barber, Jason Inman & Coral Wiggins. *Individuals exhibit idiosyncratic eye-movement behavior profiles across tasks.* Vision Research, vol. 89, pages 32 – 38, 2013.

[Rayner 98] Keith Rayner. *Eye movements in reading and information processing: 20 years of research.* Psychological bulletin, vol. 124, no. 3, page 372, 1998.

[Rayner 07] Keith Rayner, Xingshan Li, Carrick C. Williams, Kyle R. Cave & Arnold D. Well. *Eye movements during information processing tasks: Individual differences and cultural effects.* Vision Research, vol. 47, no. 21, pages 2714 – 2726, 2007.

[Rigas 12] Ioannis Rigas, George Economou & Spiros Fotopoulos. *Human eye movements as a trait for biometrical identification.* In 2012 IEEE fifth international conference on biometrics: theory, applications and systems (BTAS), pages 217–222. IEEE, 2012.

[Rigas 16] Ioannis Rigas, Oleg Komogortsev & Reza Shadmehr. *Biometric recognition via eye movements: Saccadic vigor and acceleration cues.* ACM Transactions on Applied Perception (TAP), vol. 13, no. 2, page 6, 2016.

[Salvucci 00] Dario D Salvucci & Joseph H Goldberg. *Identifying fixations and saccades in eye-tracking protocols.* In Proceedings of the 2000 symposium on Eye tracking research & applications, pages 71–78. ACM, 2000.

[Schroff 15] Florian Schroff, Dmitry Kalenichenko & James Philbin. *Facenet: A unified embedding for face recognition and clustering.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.

[Sedighi 17] Vahid Sedighi & Jessica Fridrich. *Histogram layer, moving convolutional neural networks towards feature-based steganalysis.* Electronic Imaging, vol. 2017, no. 7, pages 50–55, 2017.

[Sohn 16] Kihyuk Sohn. *Improved deep metric learning with multi-class n-pair loss objective.* In Advances in Neural Information Processing Systems, pages 1857–1865, 2016.

[Vilnis 15] Luke Vilnis & Andrew McCallum. *Word representations via Gaussian embedding*. International Conference on Learning Representations (ICLR), 2015.

[Wang 16] Zhe Wang, Hongsheng Li, Wanli Ouyang & Xiaogang Wang. *Learnable histogram: Statistical context features for deep neural networks*. In European Conference on Computer Vision, pages 246–262. Springer, 2016.

[Wu 17] Chao-Yuan Wu, R Manmatha, Alexander J Smola & Philipp Krahenbuhl. *Sampling matters in deep embedding learning*. In Proceedings of the IEEE International Conference on Computer Vision, pages 2840–2848, 2017.

[Yang 09] Ying Yang. *Penalized semiparametric density estimation*. Statistics and Computing, vol. 19, no. 4, page 355, 2009.

[Yuan 17] Yuhui Yuan, Kuiyuan Yang & Chao Zhang. *Hard-aware deeply cascaded embedding*. In Proceedings of the IEEE international conference on computer vision, pages 814–823, 2017.

[Zhang 95] Xiao Lei Zhang, Henri Begleiter, Bernice Porjesz, Wenyu Wang & Ann Litke. *Event related potentials during object recognition tasks*. Brain Research Bulletin, vol. 38, no. 6, pages 531–538, 1995.