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.

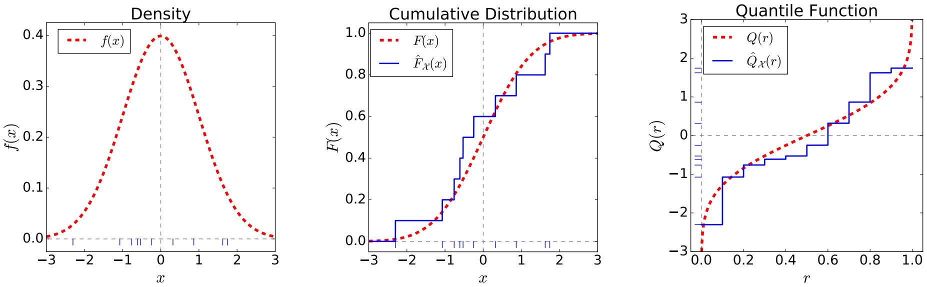

Fig. 1. Density function, cumulative distribution function, and quantile function (dashed lines) with empirical counterparts (solid lines) for a normally distributed variable $x \sim \mathcal{N}(0,1)$. Tick marks at zero line show a sample from the distribution.

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 $\hat{F}_{\mathcal{X}} : \mathbb{R} \rightarrow [0, 1]$ 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 $\hat{Q}_{\mathcal{X}} : (0, 1] \rightarrow \mathbb{R}$ is a non-parametric estimator of the quantile function Q. It is defined by

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

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 $\pi : \{1, ..., n\} \rightarrow \{1, ..., n\}$ 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

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

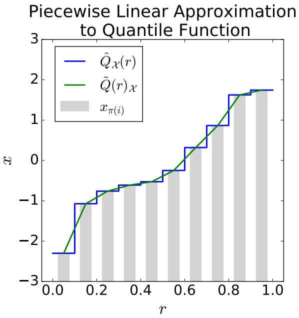

Fig. 2. Empirical quantile function, sorted samples, and piecewise linear approximation to the empirical quantile function. The set of samples is identical to that in Figure 1.

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

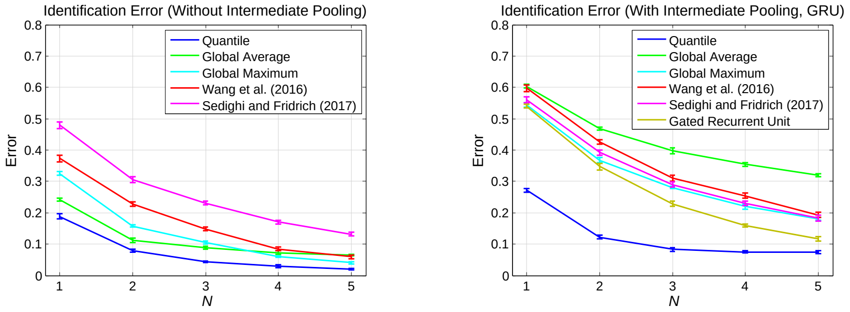

Fig. 3. Identification error for convolutional neural network architectures without intermediate pooling (left), with intermediate pooling (right) and for the recurrent neural network architecture (right) as a function of the number of test videos N on which a user is observed. Error bars indicate the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

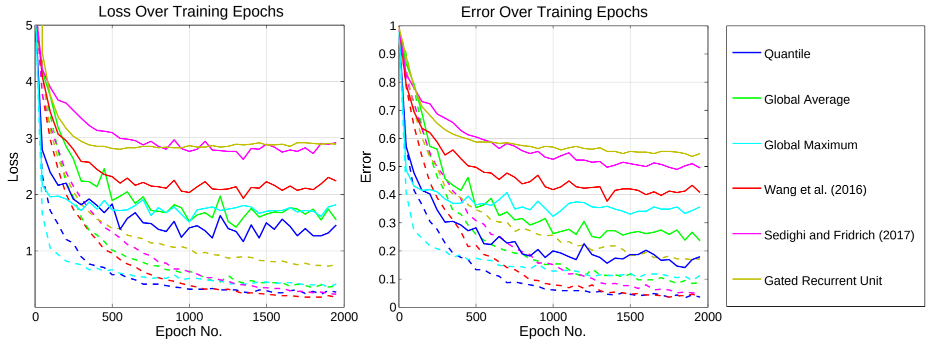

Fig. 4. Identification error (left) and loss (right) for convolutional network architectures without intermediate pooling and recurrent neural network as a function of the epoch number during training. Dashed curves denote training error and loss while solid curves denote test error and loss.

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].

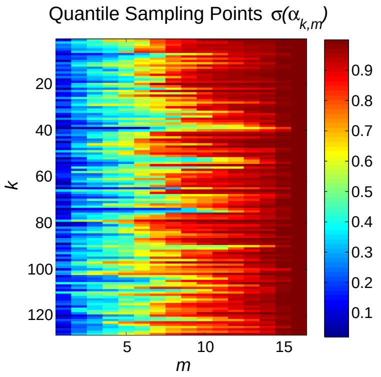

Fig. 5. Learned quantile sampling points $\sigma(\alpha_{k,m})$ as defined by Equation 10.

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.

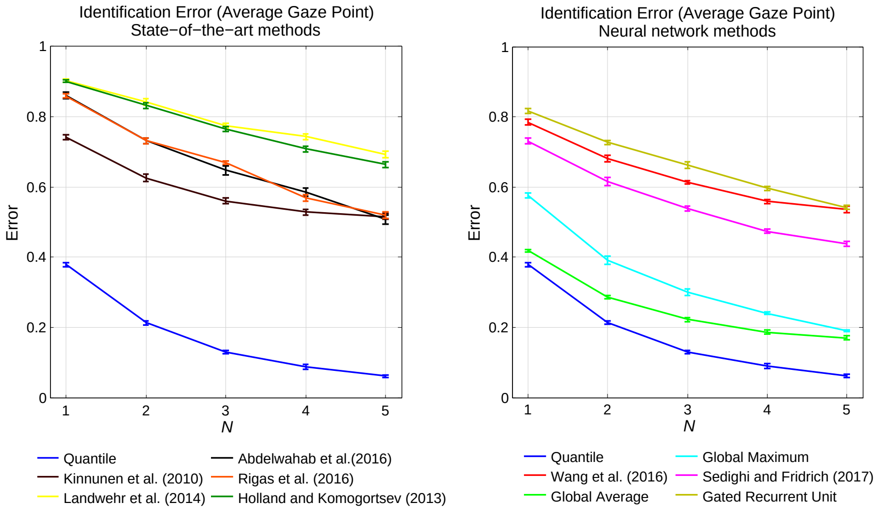

Fig. 6. Identification error as a function of the number of test videos N on which a user is observed, using average gaze point only. Error bars indicate the standard error.

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)