Paired Contrastive Feature for Highly Reliable Offline Signature Verification

Xiaotong ji[1a], Daiki Suehiro[a,b], Seiichi Uchida[a]

[a] *Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, 819-0395, Japan*
[b] *RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan*

Abstract

Signature verification requires high reliability. Especially in the writer-independent scenario with the skilled-forgery-only condition, achieving high reliability is challenging but very important. In this paper, we propose to apply two machine learning frameworks, learning with rejection and top-rank learning, to this task because they can suppress ambiguous results and thus give only reliable verification results. Since those frameworks accept a single input, we transform a pair of genuine and query signatures into a single feature vector, called Paired Contrastive Feature (PCF). PCF internally represents similarity (or discrepancy) between the two paired signatures; thus, reliable machine learning frameworks can make reliable decisions using PCF. Through experiments on three public signature datasets in the offline skilled-forgery-only writer-independent scenario, we evaluate and validate the effectiveness and reliability of the proposed models by comparing their performance with a state-of-the-art model.

Keywords: Writer-independent signature verification, Skilled forgery, Offline signature verification, Paired contrastive feature, Learning with rejection, Top-rank learning

[1] Corresponding author.
Email addresses: xiaotong.ji@human.ait.kyushu-u.ac.jp (Xiaotong ji), suehiro@ait.kyushu-u.ac.jp (Daiki Suehiro), uchida@ait.kyushu-u.ac.jp (Seiichi Uchida)

1. Introduction

The reliability of the prediction in machine learning has always been widely concerned [1]. Mainly, researchers often use class likelihood for evaluating reliability. For example, the class prediction is considered more reliable when the class likelihood of a specific class is 0.9 than 0.8. However, it is also known that the class likelihood is not accurate due to, for example, the overconfidence problem by the softmax-based classification. Consequently, specific application scenarios that require high reliability need a unique mechanism to guarantee reliability.

Signature verification [2, 3] is a typical application that requires high reliability. As shown in Figs. 1 (a) and (b), the signature verification task can be classified into writer-dependent and writer-independent scenarios[3]. For the writer-dependent scenario, we need to construct models for every individual writer. In contrast, the writer-independent scenario uses only one model that works for all writers. Accordingly, if we realize a reliable writer-independent signature verification model, it is far more practical than the writer-dependent models.

This paper proposes the *Paired Contrastive Feature* (PCF) for highly reliable writer-independent signature verification. The left part of Fig. 2 shows the process for obtaining a PCF. PCF is derived by pairing the features of two signatures; one is the genuine reference, and the other is the query signature. The features from these signatures are trained to be contrastive, enhancing the difference between genuine and forgery signatures. Thus, PCF is designed to carry the advantage of contrastive features.

The sample pairing operation of PCF realizes compatibility with machine learning frameworks for high reliability. More specifically, as shown in the right part of Fig. 2, we introduce two independent machine learning frameworks for highly reliable signature verification. The first is *learning with rejection* (LwR). Figure 3 (a) shows the general idea of LwR, where a rejection function r determines whether a sample should be rejected or not. Unlike conventional heuristic rejection rules, LwR learns a rejection function r and a rejection feature space, along with a classifier f and a classification feature space. By

[3]From another viewpoint, the signature verification task is classified as online and offline. The former deals with signatures as temporal signals and the latter as bitmaps. Although the proposed models apply to online signature verification, this paper focuses on offline signature verification.

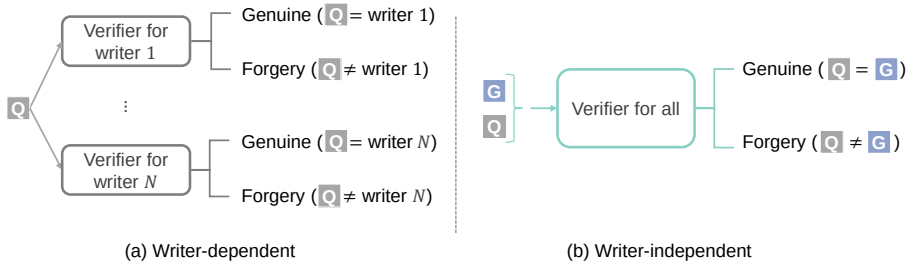


Figure 1: Two scenarios of signature verification. 'Q' and 'G' refer to query and genuine reference signatures, respectively.

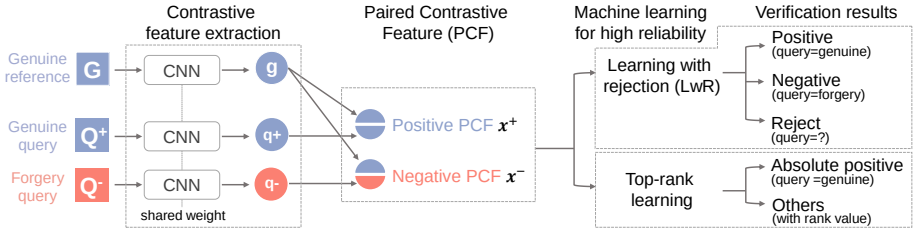


Figure 2: Overview of this paper: proposal of PCF and its application to two machine learning frameworks, *learning with rejection* and *top-rank learning*, for highly reliable signature verification.

this *co-training* process of two functions and two feature spaces, LwR realizes accurate and efficient rejection and will make more reliable predictions for non-rejected samples. As shown in Fig. 3 (b), PCF representation allows us to employ LwR for reliable writer-independent signature verification; if LwR rejects a PCF, the classifier does not predict whether the paired query and reference samples are written by the same writer or not. This applicability to LwR exhibits the usefulness of PCF, which treats a pair of samples as a single sample.

The second reliable machine learning framework is *top-rank learning*. Figure 4 (a) shows standard "learning to rank" and (b) shows top-rank learning. The former, especially bipartite ranking, provides a ranking function that gives a higher rank value to a positive sample than most negative samples; this objective is equivalent to maximizing Area Under the Curve (AUC) for the ROC [4] shown in (c). Top-rank learning (b) has a different objective; it tries to maximize the *absolute positives*, which refers to positive samples with higher rank scores than the top-ranked negative sample. By this specific objective, top-rank learning can find out "reliable" positive samples distant from all neg-

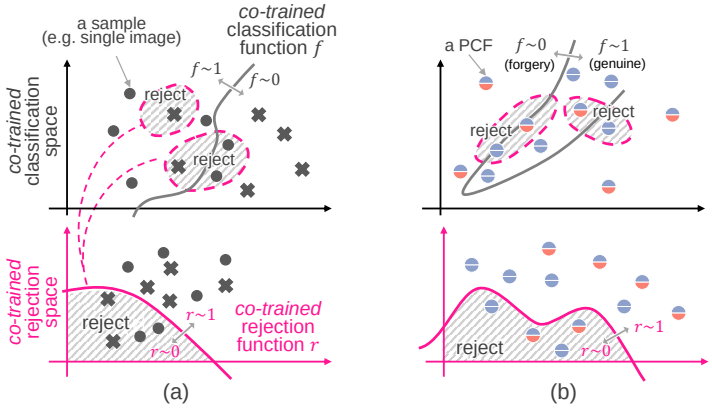


Figure 3: (a) General idea of learning with rejection (LwR), which uses a specific feature space and a rejection function for determining the samples to be rejected. (b) With PCF, LwR can handle the writer-independent signature verification task. The coordinate axis represents a two-dimensional feature space in which the data is distributed.

ative samples. Consequently, as shown in (d), by using PCF, it is possible to realize the highly reliable signature verification system that only accepts definitely genuine samples via absolute positive PCFs.

We will quantitatively and qualitatively show that the above two machine learning frameworks realize reliable writer-independent signature verification systems with PCF. For the experimental analysis, we use three public signature datasets in the "skilled-forgery-only" condition, which makes writer-independent verification far more difficult than the "random-forgery" condition. A detailed comparative study with SigNet [5], a state-of-the-art technique in the skilled-forgery-only condition, has also been conducted.

The main contributions of this work are as follows:

- We propose PCF, which converts a pairwise prediction task into a sample-wise prediction task and is thus useful for writer-independent signature verification.
- To the authors' knowledge, this is the first application of two machine-learning frameworks, LwR, and top-rank learning, to highly reliable signature verification. We emphasize that the above property of PCF makes this application possible.
- Various experiments with multiple evaluation metrics proved the high reliability of the results quantitatively and qualitatively.

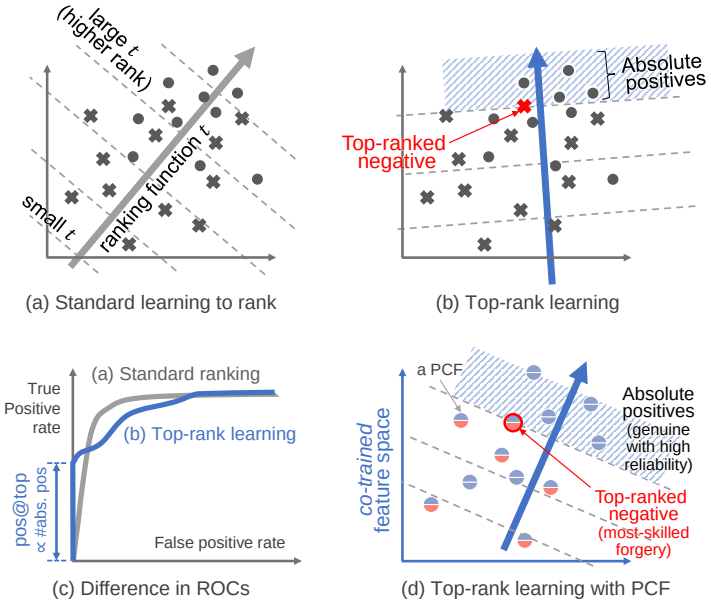


Figure 4: (a) Standard learning to rank. The ranking function is trained to give a higher rank value for positive samples (depicted as •) than negatives (×). (b) Top-rank learning, whose objective is to have more absolute positives. (c) The difference between (a) and (b) in Receiver Operating Characteristic (ROC) curves. (d) Reliable signature verification by top-rank learning with PCF.

Part of this work has been presented in a recent workshop paper [6], where the initial idea of PCF and its application to top-rank learning are introduced. From the workshop paper, this paper has significant differences in the following points. Firstly, we newly incorporate PCF into LwR for reliable verification with a rejection option. (In other words, only top-rank learning was examined in [6].) The result indicates that LwR with PCF is beneficial for reliable signature verification. Moreover, it is also shown that our LwR with PCF outperforms a rejection technique with a naive threshold operation. Secondly, we reconsider all experimental conditions and then perform all experiments from scratch to enhance the confidence of the results. For example, the results are evaluated through ten-time random (but writer-disjoint) data splittings. (In other words, there are no identical experimental results as [6].) Thirdly, we use more datasets for the proposed models in this work for a better and more general evaluation. Specifically, by newly using UTSig [7], we show that our top-rank learning with PCF can successfully detect the intrinsic difficulty of UTSig in achieving highly reliable results.

2. Related Work

2.1. Pairwise learning

Pairwise learning [8] refers to a specific learning strategy that optimizes the loss function by a pair of training samples. The loss function of pairwise learning should act on a pair instead of a single input. Metric learning [9] is a typical example of pairwise learning. The metric learning approach is committed to learning a function that gives the discrepancy between an input pair by grasping their contra distinctive feature representations. We will introduce the details in the next section.

Even though pairwise learning attracts increasing attention nowadays, there is still little effort to improve its prediction reliability. In a recent review work [10] where the pairwise learning method is one of the main topics, the authors especially introduced some "reliable" pairwise learning-based models with specific architectures and learning mechanisms [11, 12, 13]. However, even learning in a pairwise manner, existing methods treat the two inputs separately and never combine them. This is the major difference between the existing methods and the proposed PCF.

2.2. Metric learning with contrastive loss

As noted above, metric learning [9, 14] is a typical pairwise learning task. The objective of the metric learning method is to obtain an appropriate feature representation such that the samples with the same labels are close together and the samples with different labels are apart. In addition to its theoretical interests, it is useful for various applications, including signature verification[3] and person re-identification [15].

There are various metric learning methods, such as Euclidean distance[16]-based and the Mahalanobis distance[17]-based methods. Besides, [18] is a linear transformation method and [19] is a nonlinear transformation method using kernel functions [20, 21]. Further, deep learning-based methods [14, 22] with nonlinear activation functions can possess better performance than kernel-based methods for their superior feature representation capability.

Siamese network [23] is a popular deep learning-based model specialized in metric learning tasks. The Siamese network is composed of two identical networks with shared weights, which learns a representation that maps similar inputs alongside each other.

Researchers apply the Siamese network to various tasks based on its contrastive feature learning characteristic. He et al. proposed a twofold Siamese network for real-time object tracking by implementing semantic feature learning and similarity matching [24]. Ji et al. proposed a cross-attention Siamese network for video-based salient object detection by adopting a convolutional neural network (CNN)-based self-attention module and a cross-attention module along with a Siamese structure [25]. SigNet [5], also based on a Siamese network, has achieved state-of-the-art performance for the signature verification task.

2.3. Learning with rejection

An appropriate rejection operation can enhance the reliability of prediction, and it is helpful for various machine learning tasks such as handwriting recognition [26] and image classification tasks [27]. A simple and empirical rejection trial is to suspend the ambiguous predictions around the classification boundary. In an early work [28], Fumera et al. designed an algorithm for rejection based on ROC curves in binary classification. The obtained threshold is theoretically guaranteed to be optimal. Even in recent studies, the threshold-based method is still a popular choice regarding rejection [29, 30].

In contrast to the threshold-based rejection option, LwR, which learns an independent rejection function along with a classification function, is also a promising approach. The most significant advantage of LwR over its heuristic threshold-based counterpart is its ability to introduce a flexible and independent rejection function. Cortes et al. proposed a learning algorithm that can obtain a classification function and a rejection function at the same time [31] based on statistical learning theory. Mozannar et al. considered another LwR system using not only a rejection function but also the decision of a human expert [32]. SelectiveNet [33] also follows the concept of LwR and learns a highly representational classifier and rejector using a CNN-based structure. However, to the best of the authors' knowledge, the application of LwR to the signature verification task has not been fully discussed. This paper shows that the proposed PCF enables us to apply LwR to signature verification with high reliability.

2.4. Learning to rank

Learning to rank strategy [34, 35] is one of the best choices for tasks requiring high reliability such as signature verification [36, 37] and medical image recognition [38]. Among existing ranking methods, bipartite ranking [39] is a typical learning-to-rank approach. The learning objective of the bipartite ranking is to obtain a scoring function that gives higher values to positive samples than negative ones, as shown in Fig. 4 (a). This objective is equivalent to maximizing the Area Under the Curve (AUC), and thus bipartite ranking has been utilized in various domains [40, 41, 42] where high AUC is required.

Top-rank learning [43, 44] is a special type of bipartite ranking. It focuses on the ranking performance at the top rather than overall. More precisely, top-rank learning aims at maximizing the number of absolute positives, that is, to maximize the number of positive samples with higher rank scores than any negative sample. Such a learning objective is reasonable for tasks requiring highly reliable prediction like signature verification.

Zheng et al. [38] proposed a representation learning approach for top-rank learning by using Neural Network (NN) to minimize the loss of p-norm push [45]. In contrast to the previous methods based on either linear or non-linear kernel [43, 44], the model in [38] can achieve high performance even for complicated recognition tasks. While top-rank learning is appropriate for signature verification tasks, its applicability to such tasks has not yet been discussed. In this paper, we propose a novel top-rank learning model with PCF that works effectively for signature verification tasks.

2.5. Signature verification

Signature verification tasks can be categorized into writer-dependent and writer-independent scenarios [46], as explained in Section 1 and Figs. 1 (a) and (b). Also, from another point of view, signature verification tasks have online and offline scenarios [47, 48]. Online signature verification approaches use the information on signatures' pressure and stroke, and the researchers usually consider a time-series analysis. In contrast, offline signature verification uses only signature image information, which is more practical but makes verification harder. As an extensive survey, paper [2] comprehensively explores the methods and performance of signature verification models in

various scenarios, including both online and offline, as well as writer-dependent and writer-independent settings, over the last 10 years. In this paper, we concentrate on the offline writer-independent scenario according to its better practicality and greater demand for reliability.

Various methods have been proposed to obtain the discriminative features from the signature image. Zois et al. proposed a method to extract features from signatures based on asymmetric pixel relation, focusing on details of the static signature images [49]. A wrapper feature selection method is proposed in [50] to combine and select features obtained from multiple perspectives. Parcham et al. in [51] proposed a combination of CNN and Capsule Neural Networks for capturing spatial properties of signature features to improve the verification performance.

SigNet [5] is the state-of-the-art skilled-forgery-only signature verification method. It employs a Siamese network to obtain the representation for signature verification tasks via metric learning with contrastive loss. The empirical results in [5] validated the effectiveness of metric learning with contrastive loss over the standard classification approaches. Although the SigNet is proposed in 2017, it still achieves state-of-the-art performance in skilled-forgery-only signature verification tasks with a brief structure.

However, even in the writer-independent scenario, the existing signature verification methods focus on the feature mapping for each signature image. Still, they pay little attention to the contrastive relation within the signature pair. As aforementioned, in the writer-independent scenario, we have a pairwise input (query and genuine reference). So that it is more reasonable to utilize the information of a pair of signatures instead of a single signature; thus, we consider PCF, which is obtained via a contrastive feature extraction using the pairs of signature images.

3. Paired Contrastive Feature (PCF) for Signature Verification

In this section, we present the concept of Paired Contrastive Feature (PCF) for signature verification. We begin by defining PCF and explaining its components. We then discuss the utilization of PCF for signature verification tasks, emphasizing its compatibility with writer-independent scenarios. Finally, we delve into the generation method of PCFs and introduce the Siamese network as a key component.

3.1. Definition of PCF

As introduced in Section 1, PCF is a feature vector composed of two contrastive features. As shown in Fig. 2, G is a genuine reference whose real writer k is known. A genuine query Q^+ is a signature also written by the writer k. A forgery query Q^- is a signature not written by k. Let g, q^+, and q^- denote their contrastive feature vectors, respectively. We will explain the details of contrastive feature extraction in Section 3.3.

As shown in Fig. 2, a positive PCF is defined as $\boldsymbol{x}^+ = g \oplus q^+$, where $\oplus$ is the vector concatenation operation. Similarly, a negative PCF is $\boldsymbol{x}^- = g \oplus q^-$. If we have L_0 genuine references and L_g genuine queries for each of K writers in our training set, we have a positive PCF set $\boldsymbol{\Omega}^+ = \{\boldsymbol{x}_i^+\}_{i=1}^m$, where $m = L_0 L_g K$. Similarly, if we have L_f forgery query for each writer, we have a negative PCF set, $\boldsymbol{\Omega}^- = \{\boldsymbol{x}_j^-\}_{j=1}^n$, where $n = L_0 L_f K$.

3.2. Utilizing PCF for Signature Verification

The utilization of PCF enables writer-independent signature verification, and various machine-learning models can be applied. The simplest one is to use a standard binary classification model like the support vector machine (SVM); if a PCF is classified as positive, the same writer writes the query and the reference. As detailed in Section 4, this paper uses more sophisticated machine learning methods to realize a highly reliable signature verification with PCF.

The general verification process with PCF is as follows. First, we train a verification model using $\boldsymbol{\Omega} = \boldsymbol{\Omega}^+ \cup \boldsymbol{\Omega}^-$. Then, given a PCF $\boldsymbol{x} = g \oplus q$, the trained model determines whether the writer of g writes q. It is important to note that in our writer-independent scenario, the writer(s) of g and q is not included in the K writers of the training signatures. This means the trained model can apply to the PCFs of arbitrary writers.

We use PCF in the *skilled-forgery-only* condition; that is, the forgery query Q^- is written to be similar to the genuine reference G. Past attempts of signature verification often assume a more relaxed (i.e., far easier) condition, called random forgery; for example, when G is a signature of "Alice," a signature of "Bob" can be a random forgery. Therefore, it is easy to find forgeries. In contrast, our skilled-forgery-only condition does not consider random forgeries but the forgery of the same signature "Alice." Moreover,

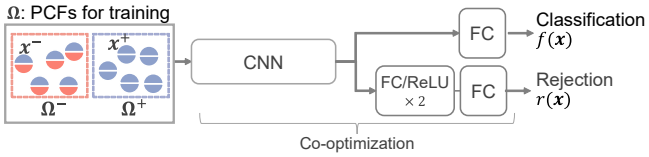


Figure 5: Reliable signature verification by LwR with PCF.

it should be *skilled*, that is, Q^- is visually similar to G. Consequently, our condition is much more challenging than the random-forgery condition. We, however, believe this challenging condition is more practical and appropriate for our "reliable" signature verification.

3.3. Generating PCFs

To generate PCF, we employ contrastive feature extraction, which aims to emphasize inter-writer differences and minimize intra-writer differences. More precisely, g and q^+ should be similar and g and q^- different. This means that we emphasize inter-writer differences and eliminate intra-writer differences before creating PCF.

We employ a Siamese network as the contrastive feature extractor for PCF. The Siamese network has also been used in SigNet [5], which is a state-of-the-art signature verification model in the skilled-forgery-only condition. The Siamese network is CNN-based and compares features of two input samples by the so-called contrastive loss. The contrastive loss becomes smaller by minimizing the distance between the feature vectors from the same class (the same writer, in our case) and maximizing the distance from different classes.

4. Two Approaches of Highly Reliable Signature Verification with PCF

4.1. Learning with rejection and PCF

As introduced in Section 1 and Fig. 3, LwR learns a rejection function and its rejection feature space simultaneously with a classification function and its classification feature space. Consequently, ambiguous samples are rejected to guarantee that the non-rejected samples are classified with high reliability. As described below, LwR has a very different framework from its heuristic counterpart, where a rejection threshold is applied to the

trained classification function. Roughly speaking, LwR follows an end-to-end training manner, whereas the latter follows a two-step manner.

Let $f(\boldsymbol{x}) \in \{0, 1\}$ and $r(\boldsymbol{x}) \in [0, 1]$ be a classification function and a rejection function, respectively. Then, we use the following decision rule:

$$(f, r)(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & \text{if } \ r(\boldsymbol{x}) \geq 0.5, \\ \text{rejected} & \text{if } r(\boldsymbol{x}) < 0.5. \end{cases} \tag{1}$$

Figure 5 shows the network structure for co-training two functions $f(\boldsymbol{x})$ and $r(\boldsymbol{x})$ with PCF. Given a PCF input $\boldsymbol{x}$, the network outputs the predictions of $f(\boldsymbol{x})$ and $r(\boldsymbol{x})$. These two functions are co-trained with the representations for $\boldsymbol{x}$ by the CNN in Fig. 5. As shown in Fig. 5, $r(\boldsymbol{x})$ employs extra fully-connected layers with ReLU. Thus, the feature space of $r(\boldsymbol{x})$ is different from the classification feature space of $f(\boldsymbol{x})$.

Following SelectiveNet [33], we use two loss functions, $\mathcal{L}_1$ and $\mathcal{L}_2$, to train the network model. The first loss function $\mathcal{L}_1$ is about rejection and is defined as

$$\mathcal{L}_1 = \mathcal{R} + \lambda \max(0, c - \mathcal{C})^2, \tag{2}$$

where λ is a hyper-parameter. Here, *rejection risk* $\mathcal{R}$ is formulated as

$$\mathcal{R} = \frac{1}{m+n} \sum_{i=1}^{m+n} l_{CE}(f(\boldsymbol{x}_i), y_i) r(\boldsymbol{x}_i) / \mathcal{C}, \tag{3}$$

where the $l_{CE}(z)$ is a cross-entropy loss and therefore $l_{CE}(f(\boldsymbol{x}_i), y_i) r(\boldsymbol{x}_i)$ is the loss for not rejecting a misrecognized sample $\boldsymbol{x}_i$. *Coverage* $\mathcal{C}$ specifies the ratio of samples *not* getting rejected:

$$\mathcal{C} = \frac{1}{m+n} \sum_{i=1}^{m+n} r(\boldsymbol{x}_i). \tag{4}$$

In the rejection risk $\mathcal{R}$, dividing by $\mathcal{C}$ is necessary to avoid a trivial solution that $r(\boldsymbol{x}_i) \equiv 0$, which implies rejecting all samples.

Eq.(2) is derived by converting the following hard-constrained minimization problem into a soft-constrained problem:

$$\min \mathcal{R} \text{ subject to } \mathcal{C} \geq c. \tag{5}$$

The hyper-parameter c specifies the ratio of samples that remains after rejection. Hereafter, we call c *target coverage* and $\mathcal{C}$ *actual coverage*. The constraint in Eq.(5) is introduced to avoid excessive rejections that make the actual coverage smaller than the

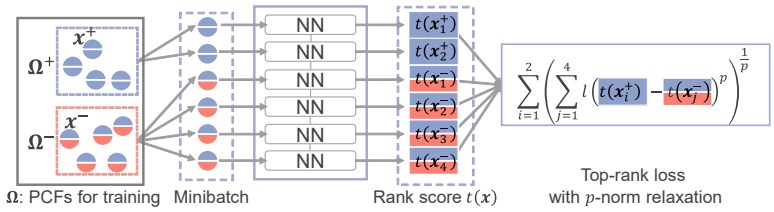


Figure 6: Training process of top-rank learning with PCF.

target coverage. The other hyper-parameter λ in Eq.(2) controls the strength of the soft-constraint term $\max(0, c - \mathcal{C})$. In Eq.(2), this term is squared for strictly penalizing the case of $c > \mathcal{C}$.

The other loss function $\mathcal{L}_2$ is introduced to optimize the classifier f independently and defined as:

$$\mathcal{L}_2 = \frac{1}{m+n} \sum_{i=1}^{m+n} l_{CE}(f(\boldsymbol{x}_i), y_i). \tag{6}$$

Finally, by combining $\mathcal{L}_1$ and $\mathcal{L}_2$, the overall loss is formulated as:

$$\mathcal{L}_{\text{overall}} = \alpha \mathcal{L}_1 + (1 - \alpha)\mathcal{L}_2, \tag{7}$$

where $\alpha \in [0, 1]$ is a hyper-parameter to balance the two loss functions.

4.2. Top-rank learning with PCF

Another framework for highly reliable signature verification proposed with PCF is top-rank learning. As noted in Section 1 and Fig. 4, the objective of top-rank learning is to maximize pos@top, which is formulated as:

$$\text{pos@top} = \frac{1}{m} \sum_{i=1}^{m} I\left(t(\boldsymbol{x}_i^+) > \max_{1 \le j \le n} t(\boldsymbol{x}_j^-) \right), \tag{8}$$

where $I(\cdot)$ is an indicator function. Pos@top refers to the ratio of "absolute positives" to all m positive samples, and the absolute positives are positive samples with higher ranking scores than *any* negative samples. More specifically, if a sample $\boldsymbol{x}_i^+$ satisfies the condition of the indicator function in Eq. (8), it is an absolute positive. The absolute positives are regarded as highly reliable positive samples. This is because they are ranked even higher than the *top-ranked negative*, which is the most "positive" negative, defined as $\max_{1 \le j \le n} t(\boldsymbol{x}_j^-)$.

Top-rank learning with PCF can realize a very strict (i.e., very reliable) signature verification. Assume that we get the ranking function $t(\boldsymbol{x})$ that maximizes pos@top. Then, we find a top-ranked negative and its rank score t_{topneg} for a given test set. Finally, if a PCF $\boldsymbol{x} = g \oplus q$ becomes an absolute positive, that is, if $t(\boldsymbol{x}) > t_{\text{topneg}}$, we verify this $\boldsymbol{x}$ to be a positive PCF and its query Q to be genuine.

This verification scheme has two important properties about the threshold t_{topneg}. First, t_{topneg} is automatically determined by the maximization process of pos@top concerning the ranking function t. Therefore, we do not need to determine the threshold by some heuristics or naive schemes. Second, t_{topneg} is specified just by a single negative sample. Accordingly, if a (very) skilled forgery can mimic the genuine signature well, this PCF with her/his forged signature will become a hard-negative with a high t_{topneg} score, making it difficult for genuine pairs being an absolute positive. In other words, it becomes difficult for genuine pairs to be accepted as genuine by the verification system[4]. This second property indicates how our verification system is reliable even in the skilled-forgery-only condition.

Figure 6 shows the training process of top-rank learning with PCF. Each PCF in a minibatch is fed to a network model for non-linear transformation and rank score calculation. Although the original objective of top-rank learning is to maximize pos@top of Eq. (8), direct maximization often fails in overfitting results. We therefore employ the p-norm relaxation technique [45, 38] for converting Eq. (8) into the following loss function:

$$\mathcal{L}_{\text{TopRank}} = \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \left(l(t(\boldsymbol{x}_i^+) - t(\boldsymbol{x}_j^-)) \right)^p \right)^{\frac{1}{p}}, \tag{9}$$

where $l(z) = \log(1+e^{-z})$ is a surrogate loss. Setting $p = \infty$, Eq. (9) is reduced to Eq. (8). Setting p at a relatively smaller value (e.g., 8 or 16), Eq. (9) casts a milder effect on the maximization of pos@top which could avoid the over-fitting issue. Another technique for stable training is the imbalance minibatch organization. During training, the positive samples in each mini-batch are ranked higher than the top negative samples *within* that

[4]A similar situation occurs when intra-writer variability is large. Positive PCFs cannot be very positive, and their rank values are not very higher than hard-negative PCFs. We will see this situation in the later experiment with UTSig.

Figure 7: Examples of genuine and forgery signatures from each dataset.

mini-batch. Therefore, each minibatch needs to include more negative samples to ensure a negative sample close to the top-ranked negative sample.

5. Experimental Results

5.1. Datasets

To evaluate the performance of the reliable writer-independent signature verification with PCF, we applied them to three commonly used signature datasets, BHSig-B, BHSig-H, and UTSig. Figure 7 shows examples of genuine and forgery from these three datasets.

BHSig-B and **BHSig-H** are two subsets in BHSig260 dataset[5] and contain Bengali and Hindi signature images, respectively. BHSig-B contains 24 genuine and 30 skilled forgery signatures for each of 100 writers. BHSig-H contains 24 genuine and 30 skilled forgery signatures for each of 160 writers. **UTSig** dataset is a Persian signature dataset collected from University students[6]. It contains 115 writers, each consisting of 27 genuine signatures, three opposite-hand signatures, six especially-skilled forgeries by experts, and 36 skilled forgeries by ordinary people. Following the past usage [7], we treated all 45 forged signatures as skilled forgery samples.

We chose these datasets because we focus on the "skilled-forgery-only" condition and therefore need to use signature datasets with a sufficient number of skilled forgery

[5] `http://www.gpds.ulpgc.es/download`
[6] `http://mlcm.ut.ac.ir/Datasets.html`

images. As noted in Section 3.2, we often find the "random-forgery" condition, where several (rather small) datasets are mixed to make random forgeries for each other in the past literature. On the other hand, in this paper, we adhere to the "skilled-forgery-only" condition because we focus on high reliability on a difficult verification task.

We used the samples from the datasets as follows. First, we prepared the same number of positive pairs (G, Q^+) and negative pairs (G, Q^-) for acquiring $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$, respectively. More specifically, in BHSig-B and BHSig-H, since we have 276 and 720 positive and negative pairs for each writer, we selected 276 randomly from 720 pairs without duplication. Then, we randomly split them into training, validation, and test sets with a ratio of $8:1:1$ by writers[7]; thus, these sets become writer-disjoint. The validation set is used for early stopping and finding proper hyper-parameters. Note that the performance was evaluated by averaging the results on ten random splittings. Each sample has been resized into 155×220 pixels before the training process.

5.2. Implementation details

5.2.1. Baseline metric leaning with contrastive loss

We employed SigNet [5] as a baseline metric learning model with contrastive loss while following the original setup. The only difference from the original setup is that we used the validation set to determine the number of training epochs. In contrast, the original setup used a fixed number of epochs. This modification was slightly beneficial for the baseline accuracy ($0.806\rightarrow 0.861$ for BHSig-B).

5.2.2. Preparing PCFs

The trained baseline, i.e., SigNet, was also used for preparing contrastive features for PCFs. Specifically, the features (g, q^+, and q^-) were extracted from the second last layer of the trained SigNet and then concatenated as positive and negative PCFs ($\boldsymbol{x}^+$ and $\boldsymbol{x}^-$), as shown in Fig. 2. The dimension of the extracted feature vector is 1,024, and therefore each PCF $\boldsymbol{x}$ is a 2,048-dimensional vector.

[7] Consequently, for BHSig-B for example, we get it $m = n = 0.8 \times 276 \times 100$ at training.

5.2.3. Preparing the LwR model

For the LwR model, we used the model of Fig. 5, where SelectiveNet [33] was combined with PCFs. As the CNN in Fig. 5, we used the VGG-16 architecture. The validation set was used in the training process to determine the hyper-parameter α among candidates {0.5, 0.6, 0.7}. Another hyper-parameter c, which controls the minimum ratio of non-rejected samples, was set at one of $\{0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0\}$. The stochastic gradient descent (SGD) was used to train the network.

We compared the performance of our model with SigNet [5] (i.e., the state-of-the-art writer-independent verification method in the skilled-forgery condition) and SigNet with a threshold-based rejection option. The latter method (denoted as "SigNet+thre") rejects a sample pair in a naive scheme. Specifically, a sample pair is rejected if the output score of SigNet falls in an interval $[\theta_1, \theta_2]$. These two thresholds, θ_1 and θ_2, are determined with the validation set. A grid search is used with two criteria. The first criterion is to make the ratio of non-rejected samples *larger than* the target coverage c[8]. The second is to achieve the highest accuracy for the non-rejected samples.

5.2.4. Preparing the top-rank learning model

As shown in Fig. 6, the inputs of the top-rank learning model are the PCFs. As for the network model (depicted as "NN" in Fig. 6), we used four fully-connected layers with ReLU. The hyper-parameter p was determined by the validation set from candidates {2, 4, 8, 16, 32}. Each minibatch contains five positive and 40 negative samples. As explained in Section 4.2, each minibatch should be imbalanced. SGD was used to train the top-rank learning network.

5.2.5. Evaluation metrics

To evaluate verification performance, we apply the following metrics to all models. **Accuracy**: the ratio of correctly classified samples. **AUC**: area under the ROC curve. **EER**: the rate when False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal. For "LwR" and "SigNet+thre," the accuracy, AUC, and EER were calculated

[8]Recall that our LwR model also expects that the actual coverage $\mathcal{C}$, i.e., the ratio of non-rejected samples, becomes larger than the target coverage c.

Table 1: The quantitative evaluation results of SigNet [5], SigNet with the rejection option ("SigNet+thre"), and our LwR model with PCFs. All values are the average of ten-time random data splittings with their standard deviation. Note that SigNet is the state-of-the-art method in the skilled-forgery-only condition.

Dataset	Methods	Accuracy (↑)	AUC (↑)	EER (↓)	Coverage $\mathcal{C}$ ($c = 0.7$)
BHSig-B	SigNet	0.861±0.040	0.935±0.031	0.137±0.042	no rejection
	SigNet+thre	0.929±0.037	0.960±0.026	0.079±0.038	0.748±0.053
	LwR	**0.945±0.036**	**0.976±0.024**	**0.061±0.047**	0.735±0.076
BHSig-H	SigNet	0.835±0.028	0.916±0.024	0.164±0.029	no rejection
	SigNet+thre	0.912±0.028	0.952±0.024	0.096±0.025	0.692±0.077
	LwR	**0.926±0.027**	**0.959±0.018**	**0.088±0.031**	0.631±0.121
UTSig	SigNet	0.670±0.014	0.744±0.021	0.323±0.017	no rejection
	SigNet+thre	0.715±0.025	0.776±0.023	**0.285±0.024**	0.757±0.029
	LwR	**0.727±0.030**	**0.778±0.038**	0.297±0.036	0.700±0.094

while ignoring the rejected samples. **pos@top**: the ratio of the absolute positives to all positives in the test set. Note that the absolute positives are defined by the top-ranked negative in the test set.

5.3. Results with LwR

5.3.1. Quantitative evaluation of LwR

Table 1 shows the quantitative evaluation results of SigNet, "SigNet+thre", and our LwR with PCFs for each dataset. In this table, we set the target coverage $c = 0.7$. For "SigNet+thre," the target coverage was set by Section 5.2.3. Table 1 confirms the following facts. First, as we expected, rejection options can increase the reliability of the verification result. By rejecting around 30% samples, LwR and "SigNet+thre" could achieve much higher accuracies than the original SigNet. Especially the positive effect of rejection is very significant in BHSig-B and BHSig-H datasets. Second, LwR achieved higher accuracies than "SigNet+thre". This fact reveals that the proposed LwR framework with PCFs can reject ambiguous samples more efficiently by utilizing its learnability

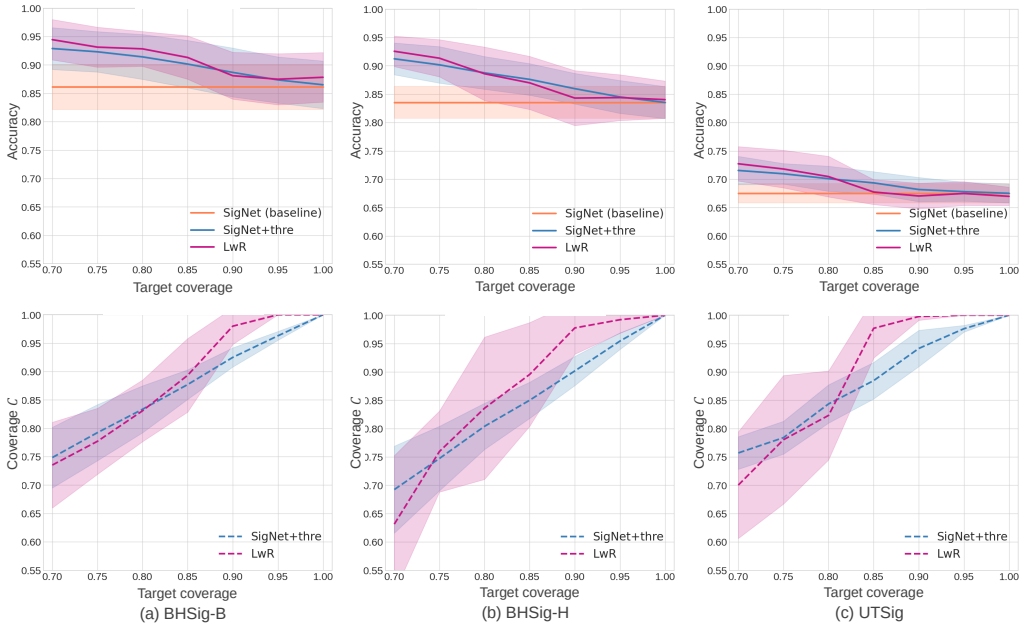


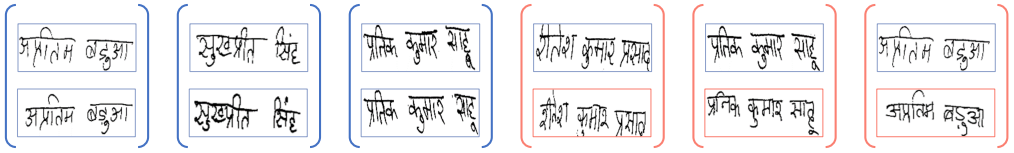
Figure 8: Average accuracy and the actual coverage $\mathcal{C}$ at different target coverage on (a) BHSig-B, (b) BHSig-H, and (c) UTSig datasets in ten-time random data splittings. The ranges within their standard deviations are filled up with lighter colors.

of the rejection function $r(\boldsymbol{x})$[9].

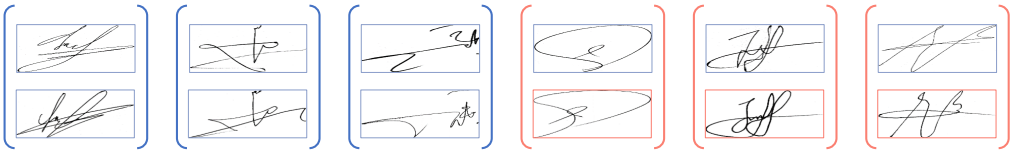
The usefulness of the rejection function $r(\boldsymbol{x})$ in our LwR model was further confirmed by other detailed observations as follows. First, $r(\boldsymbol{x})$ could reject more samples that are better to be rejected. Specifically, for BHSig-B, 61% of the misclassified samples of SigNet were rejected by $r(\boldsymbol{x})$, whereas 56% by "SigNet+thre." This indicates that our $r(\boldsymbol{x})$ could reject ambiguous samples (i.e., PCFs) appropriately to achieve more reliable verification results. The second observation is about the collaboration between $f(\boldsymbol{x})$ and $r(\boldsymbol{x})$. If we make a decision just by the classification function $f(\boldsymbol{x})$ (instead of Eq.(1)), we will have a certain amount of misclassified samples. The $r(\boldsymbol{x})$ could successfully reject

[9]To thoroughly evaluate the proposed methods, we conducted an additional experiment on dataset BHSig-B following the ICDAR2021 competition guidelines [52], where skilled forgery and random forgery signatures were simultaneously included during training and evaluated separately. The results show a high accuracy of 0.976 for skilled forgery (coverage=0.7) and a notable accuracy of 0.927 for random forgery. Importantly, there was no significant degradation by mixing skilled and random forgeries for training.

(a) Examples from BHSig-B



(b) Examples from BHSig-H



(c) Examples from UTSig

Figure 9: Examples of signature pairs that LwR rejects. Like Fig. 7, blue signatures are genuine (G or Q^+) and red are skilled-forgery (Q^-). Accordingly, the blue bracket indicates a positive pair for $\boldsymbol{x}^+$, and the red indicates a negative pair for $\boldsymbol{x}^-$. All these signature pairs are "unwanted survivors" from "SigNet+thre;" they were misclassified by SigNet but not rejected by "SigNet+thre."

around 71% of them. This reveals the effectiveness of co-training f and r.

Figures 8 (a), (b), and (c) show the accuracy and the actual coverage $\mathcal{C}$ at different target coverage $c \in \{0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0\}$, for BHSig-B, BHSig-H, and UTSig, respectively. First, the graphs show that our LwR model achieves better or equal accuracy than SigNet (the orange horizontal line) at any c. Second, when c is small ($c = 0.7 \sim 0.85$), our LwR model achieves higher accuracies than "SigNet+thre," even though their actual coverages $\mathcal{C}$ are almost the same. Interestingly, when c is large ($c = 0.85 \sim 1.0$), ours made fewer actual rejections (than c) while keeping almost the same accuracy with "SigNet+thre." (Recall that ours rejects fewer samples than specified if unnecessary.) These trends show that our rejection function r contributes appropriately to reliable verification results.

5.3.2. Qualitative evaluation of LwR

Figure 9 shows rejected pairs by our LwR. These signature pairs were misclassified by SigNet and not rejected by "SigNet+thre." Namely, they are pairs that could not

Table 2: The quantitative evaluation results of our top-rank learning with PCFs. The fact that UTSig shows a low pos@top proves the signatures in UTSig are very unstable and, thus, difficult to achieve highly reliable verification. Note again that SigNet is the state-of-the-art method in the skilled-forgery-only condition. All values are the average of ten random data splittings with their standard deviation.

Dataset	Methods	pos@top (↑)	Accuracy (↑)	AUC (↑)	EER (↓)
BHSig-B	SigNet	0.147±0.135	0.861±0.040	0.935±0.031	0.137±0.042
	Top	**0.390±0.128**	**0.882±0.042**	**0.956±0.024**	**0.112±0.041**
BHSig-H	SigNet	0.091±0.072	0.835±0.028	0.916±0.024	0.164±0.029
	Top	**0.092±0.086**	**0.838±0.034**	**0.925±0.026**	**0.153±0.032**
UTSig	SigNet	0.001±0.001	**0.670±0.014**	**0.744±0.021**	**0.323±0.017**
	Top	**0.005±0.005**	0.642±0.028	0.697±0.075	0.345±0.027

be rejected by SigNet with a threshold-based rejection option. From this figure, we can observe the following facts:

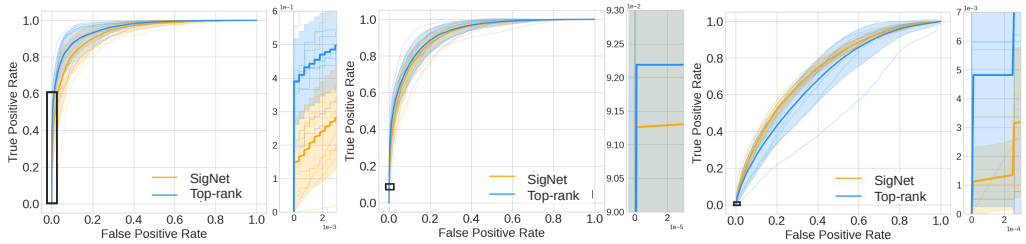
- Our LwR model rejects negative pairs $\boldsymbol{x}^-$ (pairs of genuine G and skilled-forgery Q^-, in the red brackets), appropriately, even when Q^- shows high similarity with G. Note again that "SigNet+thre" cannot reject these negative pairs. This proves the superiority of the rejection ability of LwR, whose rejection function is co-trained to reject the misclassification.

- Our LwR model could reject even positive pairs $\boldsymbol{x}^+$ (in the blue brackets) to keep the high reliability of its verification results. In other words, our model is sensitive to intra-writer differences and thus willing to reject positive pairs when the intra-writer differences are close to the differences with skilled forgeries. (Note again that these positive pairs were misclassified as negative pairs by SigNet.)

5.4. Results with top-rank learning

5.4.1. Quantitative evaluation of top-rank learning

Table 2 shows the quantitative evaluation results of the original SigNet and our top-rank learning model with PCFs. Our model with PCFs achieved higher pos@top than SigNet on all three datasets, proving that our model can determine highly reliable genuine



(a) BHSig-B dataset (b) BHSig-H dataset (c) UTSig dataset

Figure 10: The ROC curves on (a) BHSig-B, (b) BHSig-H, and (c) USTig datasets. Ten random data splittings (the thin lines) and their average (the bold line) of both SigNet and top-rank learning are shown. The ranges of standard deviations are filled up with lighter colors around the average ROC curves. The beginning parts of the ROC curves are zoomed in to observe their vertical parts that correspond to absolute positives.

signatures as absolute positives[10]. We will show examples of absolute positive samples in a later section.

Figure 10 shows ROC curves for the three datasets. The beginning part of each curve is zoomed in to observe its leftmost vertical position, whose length is relative to the number of absolute positives (i.e., relative to pos@top). For all datasets, our model shows a keener raise at the beginning of ROC than SigNet. This observation also proves that our model contributes to finding more highly reliable signatures as absolute positives.

It is noteworthy that our top-rank learning model shows better accuracy, AUC, and EER, on BHSig-B and BHSig-H. The objective of top-rank learning is to maximize the number of absolute positives, and therefore, there was a risk of degradations in the other evaluation metrics, such as accuracy. However, our model outperforms SigNet at these metrics and shows state-of-the-art performance.

[10]To ensure a fair comparison, we also evaluated a normal learning-to-rank method using ranking scores for optimization. Using the same model structure, this comparative method achieved a preliminary pos@top of 0.09. Moreover, In accordance with the training protocol outlined in the ICDAR2021 competition guidelines [52], we conducted an additional experiment on dataset BHSig-B that incorporates a combination of skilled and random forgeries during the training phase, followed by their respective evaluation. The pos@top values obtained are as follows: skilled forgery achieved 0.435, while random forgery achieved 0.005. Due to the feature extractor's specialization for skilled forgery, achieving a high pos@top in random forgery scenarios without modifying the feature extractor is challenging.

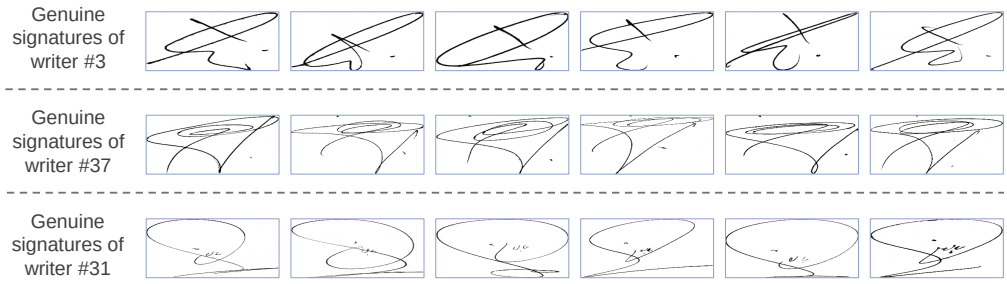


Figure 11: The reason that UTSig cannot achieve higher pos@top. It is possible to observe how the genuine signatures of three randomly-chosen writers fluctuated largely.

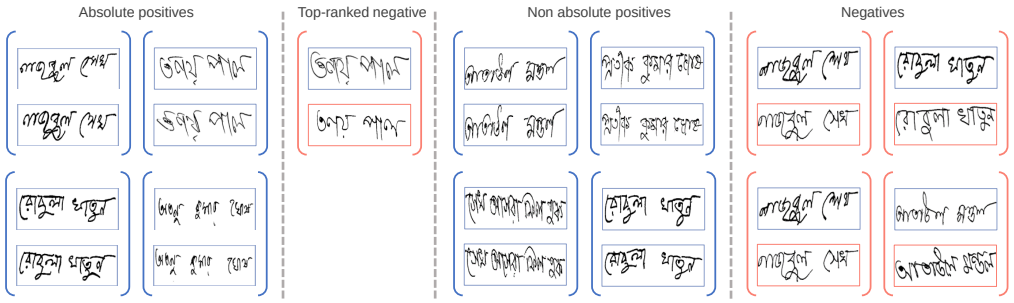
5.4.2. What does a low pos@top of UTSig mean?

The low pos@top of UTSig in Table 2 shows that our top-rank learning model can detect an unstable signature dataset that inherently has difficulty achieving high reliability. For UTSig, the value of pos@top was just 0.5% (even though it is still higher than 0.1% of SigNet). This result suggests that signatures in UTSig have significant intra-writer variability. Thus the difference between two genuine signatures is often more significant than between genuine and skilled-forgery signatures. This suggestion is confirmed by Fig. 11, where genuine signatures from three writers show significant intra-writer variability.

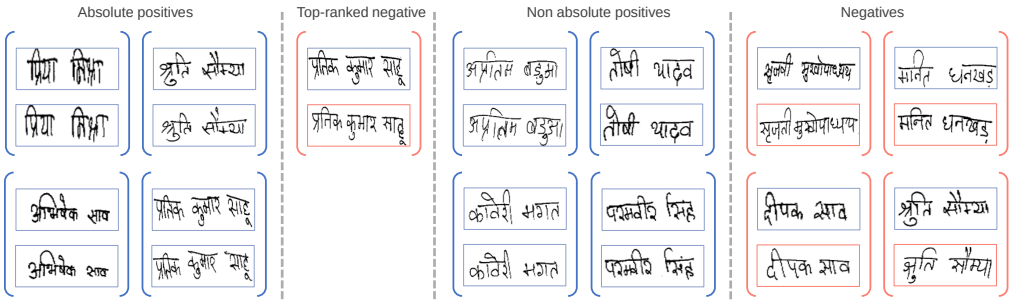
Consequently, having the low pos@top value of UTSig reveals a new aspect of signature verification. If we measure the verification performance by the standard metrics, such as AUC and accuracy, we cannot find how risky Persian signatures are in UTSig. Instead, by evaluating pos@top, we now know that UTSig has only a few absolute positive signatures, which are very hard to forge even by skilled forgery.

5.4.3. Qualitative evaluation of top-rank learning

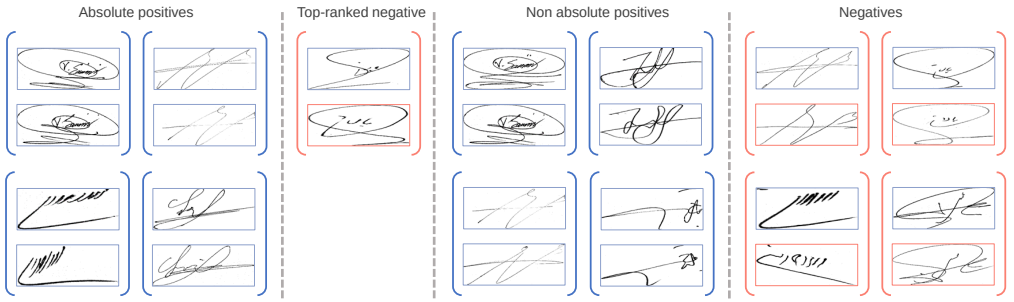
Figure 12 shows absolute positives, top-ranked negatives, non-absolute positives, and negatives for each dataset. (Except for the top-ranked negative, all those examples are randomly chosen.) This figure shows that the signature pairs in absolute positives are far more similar than other signature pairs ranked behind them. This high similarity demonstrates the high reliability of our PCF-based top-rank learning model. In our skilled-forgery-only condition, skilled-forgery signatures very similar to genuine signatures are selected as the top-ranked negatives. The absolute positive pairs achieve higher



(a) BHSig-B



(b) BHSig-H



(c) UTSig

Figure 12: Examples of the absolute positive, top-ranked negative, non-absolute positive, and negative signature pairs from three datasets ranked by our top-rank learning model.

similarity than these top-ranked negative pairs (i.e., hard-negatives) and thus are very reliable.

Comparison between absolute positive and non-absolute positive pairs also suggests that our absolute positives are reliable. Especially for BHSig-B and BHSig-H, the paired genuine samples of non-absolute positives are also very similar; they do not show large intra-writer differences. As shown in Table 2, only 39% and 9.2% of positive pairs are selected as absolute positives for BHSig-B and BHSig-H. This severe selection for absolute positives indicates that our method is useful for deriving a limited number of highly reliable signatures that will not be disturbed even by "very skilled" forgeries.

6. Limitations and Future Work

Although we verified that the PCFs work appropriately for highly reliable signature verification, the current work has the following limitations. First, we applied PCFs only to signature verification, but we can apply them to different tasks that require high reliability. For example, person reidentification can be an appropriate and direct application.

Second, various (state-of-the-art) feature extraction methods can be employed for deriving PCFs. We currently use a simple contrastive metric learning method for a fair comparison with the state-of-the-art method, i.e., SigNet [5]. We, however, can use the feature extraction and selection approach in [50] for deriving PCFs. Moreover, the framework of [51] can be the main body for the representation learning of PCFs. These combinations are expected to improve the prediction performance further.

Thirdly, Our study primarily focused on skilled forgery and did not address random forgery scenarios. This decision was motivated by our emphasis on enhancing reliability, capturing subtle features, and signature authenticity. However, future research could explore adapting our methods to tackle random forgery by incorporating specialized feature extraction techniques tailored to this specific challenge.

As part of our future work, we envision combining the proposed top-rank learning approach with the LwR method and applying them to other scenarios beyond signature verification. This integration has the potential to enhance the performance and generalizability of our methods in different application domains. Furthermore, expanding the

scope of our research to include diverse datasets is crucial. By incorporating datasets with different authenticity constraints, we can gain valuable insights into the broader applicability of our methods, evaluate their performance in real-world scenarios, and understand their effectiveness in diverse contexts.

7. Conclusion

Reliability is crucial for signature verification, especially in the writer-independent and skilled-forgery-only condition. To improve the signature verification task's reliability, we newly introduce two highly reliable machine learning frameworks, learning-with-rejection (LwR) and top-rank learning. We propose the paired contrastive feature (PCF) to use those frameworks for the task. The PCF enables handling a pair of input samples (query and reference signatures) as a single feature and thus introduces the two frameworks to the verification task.

Quantitative and qualitative experimental results verified that the proposed models (i.e., LwR with PCF and top-rank learning with PCF) could achieve reliable verification. Significantly, these methods improve the reliability of SigNet [5], which is the state-of-the-art baseline of the skilled-forgery-only verification condition. Specifically, LwR with PCF can reject samples that harm classification reliability while automatically adjusting the rejection function. The top-rank learning model with PCF is beneficial to certify highly reliable verification results as absolute positive samples, which are the positive samples ranked higher than all negative samples. The number of absolute positives also indicates the reliability of a given dataset; if the number is small, the signatures in the dataset are hard to be certified due to large intra-writer variability.

8. Acknowledgment

This work was supported by JSPS KAKENHI (Grant Number JP17H06100 and JP22H00540), Grant-in-Aid for JSPS Fellows (Grant Number JP21J21934), and ACT-X (Grant Number JPMJAX200G), Japan.

During the preparation of this work, the authors used "Grammarly" in order to correct grammatical errors. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, G. Epelde, Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing, JMIR Medical Informatics 8 (2020) e18910. doi:10.2196/18910.

[2] M. Diaz, M. A. Ferrer, D. Impedovo, M. I. Malik, G. Pirlo, R. Plamondon, A perspective analysis of handwritten signature technology, Acm Computing Surveys (Csur) 51 (6) (2019) 1–39. doi:10.1145/3274658.

[3] E. Alajrami, B. A. M. Ashqar, B. S. Abu-Nasser, A. J. Khalil, M. M. Musleh, A. M. Barhoom, S. S. Abu-Naser, Handwritten signature verification using deep learning, IJAMR 3 (2020) 39–44. doi:10.1016/j.patcog.2017.05.012.

[4] K. Uematsu, Y. Lee, On theoretically optimal ranking functions in bipartite ranking, Journal of the American Statistical Association 112 (2017) 1311–1322. doi:10.1080/01621459.2016.1215988.

[5] S. Dey, A. Dutta, J. I. Toledo, S. K. Ghosh, J. Lladós, U. Pal, Signet: Convolutional siamese network for writer independent offline signature verification, CoRR abs/1707.02131 (2017). doi:10.48550/arXiv.1707.02131.

[6] X. Ji, Y. Zheng, D. Suehiro, S. Uchida, Revealing reliable signatures by learning top-rank pairs, in: Proceedings of the DAS, 2022, pp. 323–337. doi:10.1007/978-3-031-06555-2_22.

[7] A. Soleimani, K. Fouladi, B. N. Araabi, Utsig: A persian offline signature dataset, IET Biom. 6 (2017) 1–8. doi:10.1049/iet-bmt.2015.0058.

[8] M. Huai, D. Wang, C. Miao, A. Zhang, Towards interpretation of pairwise learning, in: Proceedings of the AAAI, 2020, pp. 4166–4173. doi:10.1609/aaai.v34i04.5837.

[9] B. Kulis, Metric learning: A survey, Found. Trends Mach. Learn. 5 (2013) 287–364. doi:10.1561/2200000019.

[10] B. Lavi, I. Ullah, M. Fatan, A. Rocha, Survey on reliable deep learning-based person re-identification models: Are we there yet?, CoRR (2020). doi:10.48550/arXiv.2005.00355.

[11] E. Ahmed, M. J. Jones, T. K. Marks, An improved deep learning architecture for person re-identification, in: Proceedings of the CVPR, 2015, pp. 3908–3916.

[12] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, Joint learning of single-image and cross-image representations for person re-identification, in: Proceedings of the CVPR, 2016, pp. 1288–1296. doi:10.1109/CVPR.2015.7299016.

[13] R. R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A siamese long short-term memory architecture for human re-identification, in: Proceedings of the ECCV, 2016, pp. 135–153. doi:10.1007/978-3-319-46478-7_9.

[14] M. Kaya, H. S. Bilge, Deep metric learning: A survey, Symmetry 11 (2019) 1066. doi:10.3390/sym11091066.

[15] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. H. Hoi, Deep learning for person re-identification: A survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2022) 2872–2893. doi:10.1109/TPAMI.2021.3054775.

[16] P.-E. Danielsson, Euclidean distance mapping, Computer Graphics and image processing 14 (1980) 227–248. doi:10.1016/0146-664X(80)90054-4.

[17] R. Maesschalck, D.Jouan-Rimbaud, D.L.Massar, The mahalanobis distance, Chemometrics and intelligent laboratory systems 50 (2000) 1–18. doi:10.1016/S0169-7439(99)00047-7.

[18] M. Schultz, T. Joachims, Learning a distance metric from relative comparisons, in: Proceedings of the NIPS, 2003, pp. 41–48. doi:10.1.1.142.2314.

[19] D. Kedem, S. Tyree, K. Q. Weinberger, F. Sha, G. R. G. Lanckriet, Non-linear metric learning, in: Proceedings of the NIPS, 2012, pp. 2582–2590.

[20] A. J. Smola, B. Schölkopf, Learning with kernels, Vol. 4, Citeseer, 1998.

[21] T. Hofmann, B. Schölkopf, A. Smola, Kernel methods in machine learning, The annals of statistics 36 (2008) 1171–1220. doi:10.1214/009053607000000677.

[22] C. Shi, Z. Lv, H. Shen, L. Fang, Z. You, Improved metric learning with the cnn for very-high-resolution remote sensing image classification, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14 (2021) 631–644. doi:10.1109/JSTARS.2020.3033944.

[23] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: Proceedings of the CVPR, 2005, pp. 539–546. doi:10.1109/CVPR.2005.202.

[24] A. He, C. Luo, X. Tian, W. Zeng, A twofold siamese network for real-time object tracking, in: Proceedings of the CVPR, 2018, pp. 4834–4843. doi:10.1109/CVPR.2018.00508.

[25] Y. Ji, H. Zhang, Z. Jie, L. Ma, Q. M. J. Wu, Casnet: A cross-attention siamese network for video salient object detection, IEEE Trans. Neural Networks Learn. Syst. 32 (2021) 2676–2690. doi:10.1109/TNNLS.2020.3007534.

[26] A. Sotgiu, A. Demontis, M. Melis, B. Biggio, G. Fumera, X. Feng, F. Roli, Deep neural rejection against adversarial examples, EURASIP J. Inf. Secur. 2020 (2020) 5. doi:10.1186/s13635-020-00105-y.

[27] Z. Zhao, C. Liu, Y. Li, Y. Li, J. Wang, B. Lin, J. Li, Noise rejection for wearable ecgs using modified frequency slice wavelet transform and convolutional nural networks, IEEE Access 7 (2019) 34060–34067. doi:10.1109/ACCESS.2019.2900719.

[28] G. Fumera, F. Roli, G. Giacinto, Multiple reject thresholds for improving classification reliability, in: Proceedings of the IAPR, Vol. 1876, 2000, pp. 863–871. doi:10.1007/3-540-44522-6_89.

[29] K. Hendrickx, L. Perini, D. V. der Plas, W. Meert, J. Davis, Machine learning with a reject option: A survey, CoRR abs/2107.11277 (2021). doi:10.48550/arXiv.2107.11277.

[30] X. Yin, S. Kolouri, G. K. Rohde, GAT: generative adversarial training for adversarial example detection and robust classification, in: Proceedings of the ICLR, 2020. doi:doi.org/10.48550/arXiv.1905.11475.

[31] C. Cortes, G. DeSalvo, M. Mohri, Learning with rejection, in: Proceedings of the ALT, Vol. 9925, 2016, pp. 67–82. doi:10.1007/978-3-319-46379-7_5.

[32] H. Mozannar, D. A. Sontag, Consistent estimators for learning to defer to an expert, in: Proceedings of the ICML, Vol. 119, 2020, pp. 7076–7087. doi:10.5555/3524938.3525594.

[33] Y. Geifman, R. El-Yaniv, Selectivenet: A deep neural network with an integrated reject option, in:

Proceedings of the ICML, Vol. 97, 2019, pp. 2151–2159. doi:10.48550/arXiv.1901.09192.

[34] A. Trotman, Learning to rank, Inf. Retr. 8 (2005) 359–381. doi:10.1007/s10791-005-6991-7.

[35] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. N. Hullender, Learning to rank using gradient descent, in: Proceedings of the ICML, Vol. 119, 2005, pp. 89–96. doi:10.1145/1102351.1102363.

[36] S. Lai, L. Jin, L. Lin, Y. Zhu, H. Mao, Synsig2vec: Learning representations from synthetic dynamic signatures for real-world verification, in: Proceedings of the AAAI, 2020, pp. 735–742. doi:10.1609/aaai.v34i01.5416.

[37] Y. Zheng, Y. Zheng, W. Ohyama, D. Suehiro, S. Uchida, Ranksvm for offline signature verification, in: Preceedings of the ICDAR, 2019, pp. 928–933. doi:10.1109/ICDAR.2019.00153.

[38] Y. Zheng, Y. Zheng, D. Suehiro, S. Uchida, Top-rank convolutional neural network and its application to medical image-based diagnosis, Pattern Recognit. 120 (2021) 108138. doi:10.1016/j.patcog.2021.108138.

[39] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, D. Roth, Generalization bounds for the area under the roc curve, J. Mach. Learn. Res. 6 (2005) 393–425. doi:10.5555/1046920.1088686.

[40] N. Charoenphakdee, J. Lee, Y. Jin, D. Wanvarie, M. Sugiyama, Learning only from relevant keywords and unlabeled documents, in: Proceedings of the EMNLP-IJCNLP, 2019, pp. 3991–4000. doi:10.18653/v1/D19-1411.

[41] B. Liu, J. Chen, X. Wang, Application of learning to rank to protein remote homology detection, Bioinform. 31 (2015) 3492–3498. doi:10.1093/bioinformatics/btv413.

[42] S. Mehta, R. Pimplikar, A. Singh, L. R. Varshney, K. Visweswariah, Efficient multifaceted screening of job applicants, in: Proceedings of the EDBT/ICDT, 2013, pp. 661–671. doi:10.1145/2452376.2452453.

[43] N. Li, R. Jin, Z. Zhou, Top rank optimization in linear time, in: Proceedings of the NIPS, 2014, pp. 1502–1510. doi:10.5555/2968826.2968994.

[44] S. P. Boyd, C. Cortes, M. Mohri, A. Radovanovic, Accuracy at the top, in: Proceedings of the NIPS, 2012, pp. 962–970. doi:10.5555/2999134.2999241.

[45] C. Rudin, The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list, J. Mach. Learn. Res. 10 (2009) 2233–2271. doi:10.5555/1577069.1755861.

[46] A. Kumar, K. Bhatia, A survey on offline handwritten signature verification system using writer dependent and independent approaches, in: Proceedings of the ICACCA, 2016, pp. 1–6. doi:10.1109/ICACCAF.2016.7748998.

[47] S. Lai, L. Jin, Recurrent adaptation networks for online signature verification, IEEE Trans. Inf. Forensics Secur. 14 (2019) 1624–1637. doi:10.1109/TIFS.2018.2883152.

[48] M. Stauffer, P. Maergner, A. Fischer, K. Riesen, A survey of state of the art methods employed in the offline signature verification process, New Trends in Business Information Systems and Technology (2021) 17–30doi:10.1007/978-3-030-48332-6_2.

[49] E. N. Zois, A. Alexandridis, G. Economou, Writer independent offline signature verification based on asymmetric pixel relations and unrelated training-testing datasets, Expert Syst. Appl. 125 (2019)

14–32. doi:10.1016/j.eswa.2019.01.058.

[50] D. Banerjee, B. Chatterjee, P. Bhowal, T. Bhattacharyya, S. Malakar, R. Sarkar, A new wrapper feature selection method for language-invariant offline signature verification, Expert Syst. Appl. 186 (2021) 115756. doi:10.1016/j.eswa.2021.115756.

[51] E. Parcham, M. Ilbeygi, M. Amini, Cbcapsnet: A novel writer-independent offline signature verification model using a cnn-based architecture and capsule neural networks, Expert Syst. Appl. 185 (2021) 115649. doi:10.1016/j.eswa.2021.115649.

[52] R. Tolosana, R. Vera-Rodriguez, C. Gonzalez-Garcia, J. Fierrez, S. Rengifo, A. Morales, J. Ortega-Garcia, J. Carlos Ruiz-Garcia, S. Romero-Tapiador, J. Jiang, et al., Icdar 2021 competition on on-line signature verification, in: Preceedings of the ICDAR, 2021, pp. 723–737. doi:10.1007/978-3-030-86337-1_48.