PCA-Based Adversarial Attacks on Signature Verification Systems

Maham Jahangir[1], Azka Basit[1], Muhammad Shahzad Younis[1], and Faisal Shafait[1,2]

[1] School of Electrical Engineering and Computer Science (SEECS), National University of Sciences & Technology (NUST), Islamabad, Pakistan.
[2] Deep Learning Laboratory, National Center of Artificial Intelligence (NCAI), Islamabad, Pakistan

Abstract. Handwritten Signatures are popular biometrics that are used to authenticate individuals based on their unique physical or behavioral attributes. These systems rely on deep neural networks (DNN) for feature extraction. However, DNNs are vulnerable to small imperceptible perturbations. This research evaluates the robustness of signature verification systems through adversarial attacks. The state-of-the-art adversarial attacks possess certain limitations: first attacks being white-box in nature are impractical, and second these attacks add noise to the whole image including the background making them quite perceptible. To address these limitations, this study presents a lightweight approach based on principal component analysis (PCA). Our novel algorithm generates a universal noise vector using spatial transformation on principal components. It also strategically confines perturbation to specific regions while exploiting the principal components of the input image to launch attacks. We also computed evaluations conducted across three benchmark datasets to demonstrate compelling outcomes, in terms of attack success rate, imperceptibility, and transferability.

Keywords: Adversarial Attack · Principal Component Analysis · Signature Verification.

1 Introduction

Biometric systems are used to associate unique physical or behavioral traits to an individual which can be used to authenticate and identify these individuals mostly in legal and financial matters. Since these are unique features it's hard to replicate or forge them. The handwritten signature is one such biometric trait that has been used for centuries to authenticate or verify various entities related to humans like bank cheques, documents, forms, and many more. The Signatures are an important biometric because their collection is widely acceptable, smooth, and non-invasive. These are modeled as pattern recognition and machine learning systems where the signatures of a user are stored in the systems as a reference or template and the signature under test is authenticated to be either original or forged. These systems can be online (dynamic) or offline (static).

In an online system, the signature is acquired at the runtime with a stylus and digital pad, and attributes like velocity, pressure, stroke length, writing speed, etc. are calculated. Online systems are more powerful and accurate but are quite expensive as compared to offline systems. Moreover, offline-based systems are inevitable in certain situations like cheque transactions, etc. In this research, we have considered the case of offline signature verification systems. Another categorization of these systems is based on the writer. One is writer-dependent and the other is writer-independent systems. The writer-dependent systems are updated every time a new user gets registered in the system. On the contrary, the writer-independent systems don't follow that constraint and are generally considered more practical [8]. The scope of this research article covers writer-independent systems.

In the past, signature verification systems relied on handcrafted features, but since the resurgence of Deep Neural Networks (DNN)-based systems, there has been a paradigm shift toward more automated and data-driven approaches in the field. Like all other areas, DNN has achieved state-of-the-art performance in signature verification. The popularity also comes with the need to evaluate the robustness of these systems. DNNs are said to be vulnerable to small carefully crafted perturbations/noises [7]. These noises are added to the input image. These images are then called adversarial examples. The adversarial examples fool the classifier misleading it into predicting wrong labels. The attack process is called an adversarial attack. In this research article, we evaluated the robustness of the Signature Verification Systems against our proposed adversarial attack method as well as against state-of-the-art methods.

Since the advent of adversarial attacks by Szedgy et al. [17] a lot of attacks have been designed and proposed by various researchers. Some of the state-of-the-art include FGSM [7], Carlini and Wagner [3], BIM [12], etc. All these attacks have been mostly designed for classification systems. It is to be noted that attacking a verification system is not the same as attacking other classification systems. The former possesses challenges and limitations that require careful consideration. First of all, every time a new user enters systems an unseen class is introduced. Secondly, the signature images are just strokes of the signature on a bare background. Most of the attacks are applied to the full image but in the signature images these attacks don't provide good results as they attack the full background and these pixels are already not taken into account by the signature verification systems. Moreover, most of the attacks introduced in the literature are white-box in nature i.e. they require full information of the model to craft the attack. This is quite impractical in the case of signature verification systems since these systems are involved in various legal and financial matters so access to these or the information of the model is quite well protected. We are interested in exploring whether the area of verification systems possesses any characteristics that make them vulnerable to adversarial attacks as well. In the light of these problems highlighted we present a novel attack method based on exploiting the principal components of an image. Our approach doesn't need any information regarding the DNN model employed by the system. It's a model-free

approach. Moreover, with the use of principal components of an image we try to alter selective pixels of the image to create an attack. The main objective is to create a black-box, transferable attack to evaluate the robustness of signature verification systems.
Principal Component Analysis (PCA) is essentially a data dimensionality reduction approach. It represents data linearly by reducing its dimensionality but retaining important information. PCA transforms the original data into a new data manifold where the new variables are uncorrelated known as principal components, but capture the maximum variance in the data. It is a very lightweight linear approach to transform the data into new space. PCA is utilized for extracting meaningful features from complex datasets, aiding in the identification of critical variables or patterns [6].
Adversarial examples are regarded as non-robust features of the input images [17]. Motivated by this concept we make use of principal component analysis to highlight the non-robust components that correspond to non-robust features of the input images. We add noise to these components instead of the full image to create an attack. The key here is to add noise to the minimum but important components of the image. We generated universal noise, using spatial transformation on principal components. We have carried out extensive experimentation in this area to evaluate the strength of the proposed attack method. In short in this research article, we make the use of principal component analysis to find out the minimum set of components to be altered in a way that the difference is minimal but big enough for the model to be fooled. We conducted experiments on three benchmark datasets trained on Siamese Convolutional Networks.

The main contributions and findings include:

1. The proposed model can generate imperceptible adversarial examples without requiring gradients or model architecture making them practical to deploy.
2. We developed two new techniques regarding adversarial attacks on signature verification networks using a lightweight phenomenon Principal Component Analysis (PCA). One is the restriction region of perturbation applied on principal components of the image as well as creating a universal perturbation matrix generated by spatial transformation on principal components themselves.
3. We also evaluated the transferability of the proposed approach making the attack more practical and hence difficult to defend.

The structure of the paper is as follows. Section 2 highlights the limited amount of existing literature on the topic under consideration. Next, we discuss the detailed methodology in Section 3. Experimental Protocol is described in Section 4, followed by the results and discussion in Section 5. Finally, the paper is concluded in Section 6.

2 Related Work

Adversarial examples are the input images that are intentionally perturbed with noise that is used to fool the classifier into wrong prediction. Since the advent of adversarial attacks by Szegedy et al. [17] in 2013, a lot of attacks and defense mechanisms have been proposed by researchers and practitioners to evaluate the robustness of deep neural networks. Some of the first-generation state-of-the-art methods include Fast Gradient Sign Method (FGSM) [7], Basic Iterative Method (BIM) [12], Projected Gradient Method (PGD) [14] and many more. Universal adversarial attacks are also introduced by Moosavi et al. [15] in which a single noise is used to perturb all the data input belonging to different classes. FGSM is a white box attack method that uses gradients to maximize the loss of the classifier to perturb the input images. BIM is an extension of FGSM which is iterative and takes small steps toward maximizing the loss of the classifier. PGD is a sophisticated extension of BIM which is an iterative optimization-based method. Similar to BIM, PGD starts with an initial guess of the adversarial perturbation and iteratively updates it in the direction that maximizes the loss function, while ensuring that the perturbation remains within a bounded epsilon-ball around the original input. All these attack methods are white box in nature and require full information of the model and gradients to craft the almost impossible attack, especially in sensitive secure systems like signature verification.

Adversarial attacks against signature verification systems are a relatively underexplored area with only a handful of research articles in the area. Hafemann [9] evaluated the robustness of signature verification systems using existing adversarial attack methods like FGSM [7] and C&W [3]. These methods are quite perceptible which disrupts the background of the signature image. Most importantly all the methods require full information about the system the gradients, and the network architecture which is not practical. In another research Li et al [13] proposed a black box method with region restriction. It's an iterative approach where noise is optimized while restricting it to the region of strokes. The main drawback of this approach does not apply to binary images because pixel values are not continuously adjustable and cannot select optimal pixels. In another research [10] dictionary learning-based approach is used to attack the signature verification systems. Our research is inspired by the same research with the primary difference in the selection of data representation technique. They [10] used dictionary learning to learn sparse representations of data and use these representations as noise to create adversarial attacks. We have used a lightweight approach to achieve the same with state-of-the-art performance, whereas dictionary learning is computationally expensive. Moreover, they did limited experimentation whereas our research did extensive experimentation to prove the efficacy of the proposed approach. Only these three articles evaluated the signature verification systems against adversarial attacks to the best of our knowledge.

Another area of related work is the use of Principal Component Analysis (PCA) to create adversarial attacks. This approach has been used by the authors in [21]. They evaluated their proposed approach principal component adversarial

example (PCAE) on MNIST, CIFAR, and Imagenet datasets where they didn't receive very encouraging results, especially on grayscale images like MNIST. We also conducted experiments to compare our proposed method with PCAE. Principal Component Analysis is also used to attack the audio domain of data by authors in [1]. In this research, we have used PCA to compute a universal noise to exploit a specific region of the principal components of the input image. Since the background and foreground are separated, attacking signature verification systems is very hard and challenging when compared to other classification systems. We attacked a Siamese network as they are gaining fast-growing popularity due to their remarkable results [4, 18]. We have also highlighted the shortcomings of traditional region restriction approaches in detail in section 3.3 followed by the detail on how our method of region restriction is better. The experimental and evaluation section proves the efficacy of our proposed approach.

3 Our Approach

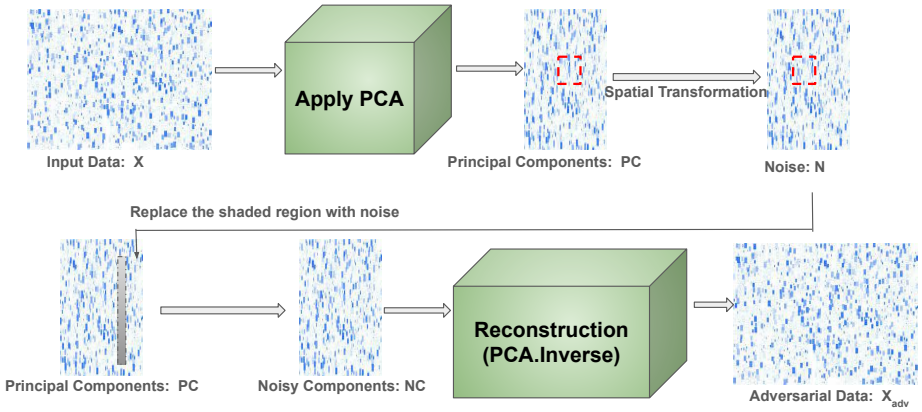

Fig. 1. Framework for the PCA-based Proposed Attack Model. The first row illustrates how universal noise is computed and the second row shows exploiting specific principal components of the image with noise to create adversarial examples.

We explain our proposed approach in detail in this section as illustrated in Figure 1. The first step in the proposed approach is to compute the Principal Components of the input image. On the other hand, a universal noise vector is also computed using spatial transformation and principal component analysis. In the third step, this noise is added to the principal components of the original input image. Next, the noisy and non-noisy components are combined to reconstruct the image. The reconstructed image is our adversarial image which is then fed to the Siamese network for evaluation. In this section, all these phases

are discussed in detail along with the problem definition and the details of the Siamese Network.

3.1 Problem Definition

In this section, we formally define the problem and explain the threat model under consideration. The model used in our research is the Convolutional Siamese Network. Siamese networks have achieved remarkable results for the signature verification problem. These are twin CNN architectures that share the same parameters and learn the same weights. These learn the same feature space when introduced with similar and dissimilar inputs. This is achieved usually by minimizing Euclidean distance between similar pairs of inputs and maximizing it between dissimilar pairs. These are ideal for problems involving similarity comparisons between pairs of data. As in this case similarity between original and forged images. The loss function usually used for these network architectures is contrastive loss which is defined as follows:

$$L(x_1, x_2, y) = \alpha(1-y)D_w^2 + \beta y \max(0, m - D_w)^2 \quad \text{where} \quad x_1, x_2 \in X \quad (1)$$

Here x_1 and x_2 are input samples that can be original or forged signatures. $D_w = \| f(x_1; w_1) - f(x_2; w_2) \|_2$ is the Euclidean distance and w_1 and w_2, are learned weights. y is the binary class label that denotes whether the two input samples are similar or dissimilar. The forgers try to fool the signature verification systems by creating forgeries of the signatures of the user. The DNN-based systems effectively detect these forgeries as the Siamese Network used in this paper achieves 100% accuracy in detecting original and forged signatures. However, such systems still suffer from two main threats identified as adversarial attacks in this paper. One of the input samples introduced to the model is poisoned with a small perturbation denoted as x_{adv} which fools the model into assigning a wrong indicator to the input pair. Mathematically, in the case of Siamese networks, it can be denoted as:

$$L(x_1, x_{\text{adv}}) \neq y \quad (2)$$

where,

$$x_{\text{adv}} = x_2 + \epsilon p \quad \text{and} \quad d(x_{\text{adv}}, x_2) < \epsilon \quad (3)$$

The signature verification networks can suffer from two types of threats: Type I and Type II attacks. Type I also known as False Rejection means that the genuine signatures are modified in a way that they are rejected by the system. This isn't a very practical scenario as one needs to have access to the original signatures of the user. The second is Type II also known as False Acceptance. In this case, the forgeries are modified in a way that they are accepted by the system. Type I requires access to original signatures by users which again is not a very practical option for such secure systems. Type II however is more practical where you modify a forgery to be accepted. In this article, we performed experiments for type II attacks that is False Acceptance.

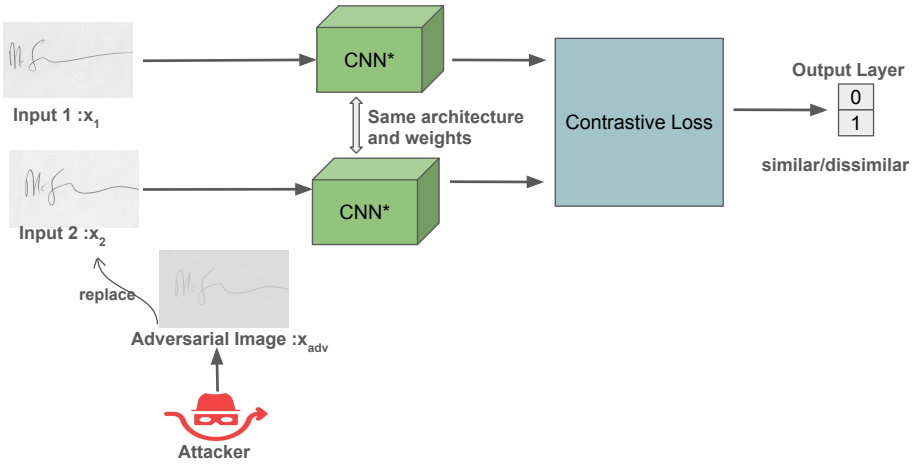

Fig. 2. Siamese Convolutional Network-based Signature Verification System under adversarial attack threat model

3.2 Principal Components

Principal Component Analysis is a statistical measure that is used to represent data with linear combinations and lower dimensions. This tool transforms the data into a new representation of uncorrelated data points with maximum variability. The new data components are sorted according to the decreasing amount of variance with the first component holding the maximum variance and second holding the second maximum variance and so on. Mathematically, PCA aims to calculate the orthogonal axis (principal components) along which data varies most. It does so by computing the covariance matrix on standardized data.

$$\Sigma = \frac{1}{n-1}(X' - \bar{X}')^T(X' - \bar{X}') \tag{4}$$

X' is the standardized data matrix and $\bar{X}'$ is the mean vector of the standardized data.
The next step is to perform eigenvalue decomposition on the covariance matrix to compute eigenvectors (V) and eigenvalues (λ). Next, we select the k principal components that are k largest eigenvalues to form V_k

$$PC_i = X'v_i \tag{5}$$

where PC_i is the i-th principal component and v_i is the i-th eigenvector.
PCA has proven to be very helpful in explaining the behavior of neural networks. Since it reduces data dimensionality while retaining the most important information with maximum variance, thus helps in the interpretability of the model [20]. The feature reduction helps identify important features. Therefore, the motivation behind this research article is to make use of this important information to create adversarial attacks. We have used PCA for two tasks. First

to compute a universal noise via spatial transformation on principal components and secondly, we add the said noise to the principal components of the image rather than the image itself. These two phases of the methodology are explained in the preceding sections 3.3 and 3.4.

3.3 Restricting Perturbation to Specific Regions

The signature images typically have a clear background, separating it into foreground and background elements, which complicates attempts at adversarial attacks. Previous studies, such as those outlined in [13] and [10], employed conventional methods like image inversion, GrabCut segmentation, or setting background pixels to 0 to delineate foreground from background. However, these methods have limitations. For instance, setting the background to 0 and the foreground to 1 is not viable for binary images. Moreover, most of the signature verification models already extract the foreground as part of their pre-processing step.

In our research, we introduce a novel technique aimed at confining noise perturbations to specific regions of the image, targeting critical features essential for model classification. To address existing limitations, our study presents a lightweight approach rooted in Principal Component Analysis (PCA). Our novel algorithm strategically limits perturbations to specific regions while exploiting the principal components of the input image to execute attacks. Notably, our attack method is agnostic to model specifics, requiring no knowledge of the model, its architecture, or gradients.

The process begins by transforming the image into a new feature space using PCA, yielding a list of components containing the most important information. The components hold maximum variance among data. Subsequently, noise is selectively applied to regions within each component, preserving imperceptibility by targeting only specific areas rather than the entire image.

Mathematically already defined in 5 this can be represented as follows:

$$PC = X'V_k \tag{6}$$

where PC stands for Principal Components computed after applying PCA to data (X) and V_k are first k eigenvectors.

$$nc_i = PC_i + region \times noise[i] \tag{7}$$

where nc_i is the i-th noisy component. Next, we explain the computation of this noise to be added in the next section.

3.4 Universal Noise Computation via Spatial Transformation on Principal Components

In this research, we have used the principal components themselves to create a universal noise matrix. The principal components of X are calculated followed by

spatial transformation operation. Spatial Transformation refers to altering the spatial arrangement of the pixels in an image. These alterations are geometric like rotation, and translation scaling that is changing the spatial relationship between pixels of an image. The main idea is to tamper the components of the image with a noise that holds relevant information about the data. The key is to control the intensity and scale of noise in the important region which will ensure its imperceptibility, transferability, and effectiveness against defense. We have conducted experiments to evaluate all three parameters. Spatial Transformation has been used by researchers to craft adversarial attacks and shows significant strength [19]. In our case, we compute the principal components of the data. Next, The rotation is performed around the center of the components matrix, which is specified as $a_c, b_c = (cols/2, rows/2)$. The rotation angle θ is set to 180 degrees, indicating a full 180-degree rotation. This transformation is applied to the components (PC) and stored as $noise$ variable.

$$\begin{bmatrix} a' \\ b' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & (1-\cos(\theta)) \cdot a_c + \sin(\theta) \cdot b_c \\ \sin(\theta) & \cos(\theta) & -\sin(\theta) \cdot a_c + (1-\cos(\theta)) \cdot b_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ 1 \end{bmatrix} \tag{8}$$

Where, (a, b) represents the coordinates of each pixel from the components computed i.e. $(a, b \in PC)$, $(a^{'}, b^{'})$ represents the coordinates of the corresponding pixel in the rotated matrix, θ represents the rotation angle, and (a_c, b_c)represents the center of rotation.

3.5 Reconstruction from Principal Components to generate Adversarial Images

As a final step after the noise is added to specific regions in the principal components of the image the original image is reconstructed using the PCA inverse transform function. The reconstructed image is the adversarial image with embedded noise in its components. This adversarial image is then fed to the Siamese Convolutional Neural Network and is given as

$$X_{adv} = PCA.inversetransform(NC) \tag{9}$$

where X_{adv} represents the set with all the adversarial images generated and NC denotes the Noisy Components of all the images. $NC = nc_1, nc_2, ...nc_k$ computed in equation 7.

4 Experimental Setup

The experimental protocol designed to evaluate the proposed approach is discussed in this section. The details on datasets, architecture used, baseline methods, and metrics used for evaluation are all listed in this section.

Algorithm 1: Principal Components based Adversarial Examples

Input: $X \rightarrow$ Set of pre-processed original signature images;
Result: $X_{adv} \rightarrow$ Adversarial Examples Set

1. Apply PCA to X
$PC = PCA.transform(X)$

2. Generate noise (N)
$N = SpatialTransformation.Rotation(PC)$

3. Generate noisy components (NC)
$nc_i = PC_i + region \times N[i]$
$where,\ \ nc_i \in NC$ set

4. Generate adversarial examples (X_{adv})
$X_{adv} = PCA.inversetransform(NC)$

Return X_{adv})

Table 1. The details of datasets. Number of users and original and forged signatures per user used for the experimentation.

Dataset	Number of users	Genuine images per user	Forged images per user
Cedar	55	24	24
MCYT	75	15	15
GPDS	300	24	24

4.1 Datasets

We conducted experiments on three popular and widely used benchmark datasets: CEDAR [11], MCYT-75 [16] and GPDS-synthetic [5]. The dataset contains original signature images by legitimate users as well as forged signature images. The details on the number of users and the number of original and forged signature images per user are tabulated in Table 1. We have used all users' data for Cedar and MCYT but for GPDS we used 300 users with 24 forged and original images per user.

4.2 Network Architecture

The proposed method is evaluated on a Siamese-based Convolutional Neural Network as outlined in the paper in the paper [4]. The pre-processing steps include resizing the image to a fixed size of (155×220) followed by an inversion operation to extract the foreground with a black background and white foreground. We used the publicly available implementation of model architecture with a minor variation in one of the layers [3]. The variation includes changing the filter size in one of the convolutional layers. This model is named Signet by

[3] https://github.com/AtharvaKalsekar/SigNet/

the authors and it has achieved state-of-the-art performance in signature verification systems. Therefore, its powerful feature extraction makes it hard to attack it.

4.3 Metrics

The contrastive loss, attack success rate, and structural similarity index measure (SSIM) are measured to evaluate the proposed approach. The attack success rate defines the number of forged signatures that are declared original by the system (False Acceptance). SSIM is used to measure the similarity between two images. It evaluates three aspects of similarity: luminance, contrast, and structure. SSIM is widely used in image processing and computer vision tasks to assess the quality of compressed or distorted images. Its value ranges from 0 to 1, where 0 means no similarity and 1 means full similarity. SSIM is given as:

Fig. 3. From left to right: First image is the legitimate forged input image with no noise added, Second is the sample adversarial image generated by the proposed approach followed by the adversarial images generated by state-of-the-art methods.

4.4 Baseline Methods and other Noises

We ran experiments to compare the proposed methods with two categories of attacks. One is the baseline method. Fast Gradient Sign Method (FGSM) [7], Basic Iteration Method (BIM) [12], Projected Gradient Descent [14], and Boundary Attack [2] are among the first-generation state-of-the-art baseline methods. We also compared the proposed method with the only relevant method related to principal components to the best of our knowledge. This method is listed in the tables by the name Principal Component Adversarial Example (PCAE) [21]. The second set of experiments conducted was to compare the proposed method with a different set of noises. We ran experiments for Gaussian, Speckle, Salt & Pepper, and Spatial Transformation noises.

5 Results and Discussion

This section discusses the results achieved when the proposed method is applied to Cedar, MCYT, and GPDS datasets and compared with other methods. Moreover, discussion on different factors affecting the performance of the methodology is also discussed with reference sample images.

Table 2. The Loss of Classifier (lower the value more successful the attack is), Attack Success Rate, SSIM values reported on Cedar Dataset for our proposed method, different noises, and baseline methods.

Method	Loss	Attack Succ. (%)	Mean SSIM	Median SSIM
Baseline Methods				
FGSM [7]	0.07	88%	0.3	0.3
BIM [12]	0.07	88%	0.37	0.36
PGD [14]	0.07	88%	0.36	0.36
Boundary Attack [2]	2.74	66%	0.48	0.34
PCAE [21]	0.09	87%	0.76	0.72
Different Noises				
Gaussian	10.50	0.5%	0.97	0.98
Speckle	11.70	0%	0.95	0.95
Salt & Pepper	0.09	88%	0.04	0.04
Spatial Transformation	0.08	89%	0.48	0.48
Proposed Method	**0.07**	**90%**	**0.71**	**0.73**

Table 3. The Loss of Classifier (lower the value more successful the attack is), Attack Success Rate, SSIM values reported on MCYT Dataset for our proposed method, different noises and baseline methods.

Method	Loss	Attack Succ. (%)	Mean SSIM	Median SSIM
Baseline Methods				
FGSM [7]	0.4	23%	0.41	0.41
BIM [12]	0.4	38%	0.47	0.47
PGD [14]	0.4	36%	0.47	0.47
Boundary Attack [2]	1.5	16%	0.7	0.9
PCAE [21]	1.86	19%	0.9	0.9
Different Noises				
Gaussian	1.88	24%	0.95	0.98
Speckle	1.2	25%	0.9	0.97
Salt & Pepper	0.4	51%	0.04	0.03
Spatial Transformation	0.22	70%	0.38	0.37
Proposed Method	**0.3**	**71%**	**0.71**	**0.71**

5.1 Results

We have reported contrastive loss, attack success rate, mean and median SSIM values for the proposed approach, state-of-the-art, and different noises on the principal components of the image. Table 2 reports all these metrics for the experiments conducted on the Cedar Dataset. The results show the proposed method achieves the highest attack success rate with the lowest loss. The SSIM value for the proposed approach is 0.7. The state-of-the-art gives almost similar results with very low values of SSIM i.e. 0.3. Fig 3 shows how state-of-the-art approaches like FGSM, and BIM disrupt the whole image making noise perceptible. On the other hand, the PCAE [21] method gives promising results highlighting the efficacy of principal components-based adversarial attacks. The results tabulated in Table 3 show that the proposed method achieves the highest success rate among others with an SSIM value of 0.7 when experiments are conducted on the MCYT dataset. In the case of GPDS results reported in Table 4 show the proposed method only achieves an attack success rate of 30%. Although it is higher among other methods this low success rate is because data is synthetic.

Table 4. The Loss of Classifier (lower the value more successful the attack is), Attack Success Rate, SSIM values reported on GPDS Dataset for our proposed method, different noises, and baseline methods.

Method	**Loss**	**Attack Succ. (%)**	**Mean SSIM**	**Median SSIM**
Baseline Methods				
FGSM [7]	0.7	27%	0.8	0.8
BIM [12]	0.4	28%	0.67	0.67
PGD [14]	0.49	28%	0.67	0.68
Boundary Attack [2]	0.4	16%	0.8	0.8
PCAE [21]	0.70	2%	0.67	0.67
Different Noises				
Gaussian	0.79	19%	0.9	0.9
Speckle	0.43	15%	0.5	0.6
Salt & Pepper	0.82	3%	0.2	0.2
Spatial Transformation	1.55	3 %	0.38	0.36
Proposed Method	**0.49**	**30%**	**0.75**	**0.76**

5.2 Effect of Different Noises

We have reported results of Gaussian, Speckle, Salt & Pepper, and Spatial Transformation noise applied on principal components of the image on Cedar, MCYT, and GPDS datasets in Tables 2, 3 and 4 respectively. The results for Gaussian and Speckle noise are not encouraging. They don't attack the system at all. Salt and Pepper and spatial transformation noise achieve better attack success rate but with no imperceptibility. The images from Fig 4 illustrated how the image

is completely distorted in the case of Salt & Pepper as well as Spatial Transformation. The high attack rate of spatial transformation noise encouraged the proposed method to apply the spatial transformation-based universal noise to specific regions of principal components to improve the imperceptibility with a high success rate and low loss.

5.3 Transferability

The proposed method is considered a strong attack method as it's a black-box attack also known as gradient gradient-free method. Another strength of the proposed method is that it is data-free as well. We evaluated the transferability of the proposed method across different datasets. The results tabulated in Table 5 show the results when one dataset is considered a source and noise generated from the source dataset is applied to the target dataset and model according to the proposed methodology. We have achieved high success rates i.e. 90%, 77%, and 56%.

Fig. 4. Adversarial image samples of different noises applied to principal components of the image on Cedar, MCYT, and GPDS datasets

5.4 Effect of Region Restriction

The backbone of the proposed approach is restricting perturbation to specific regions. Fig 5 illustrated the effect of different regions on the attack success rate

and SSIM. We selected different regions of the principal component vector and perturbed it with universal noise. It is evident from Figure 5 that when noise is added to components in the beginning that hold maximum variance the effect on imperceptibility is the greatest. As we move towards the components that hold less variance the imperceptibility increases.

Table 5. The Loss of Classifer, Attack Success Rate, and SSIM reported for the case of transferability with one dataset as the source and other as target

	Cedar			**MCYT**			**GPDS**		
Source	**Loss**	**Attack Succ.**	**SSIM**	**Loss**	**Attack Succ.**	**SSIM**	**Loss**	**Attack Succ.**	**SSIM**
Cedar	-	-	-	0.81	54%	0.8	0.56	24%	0.8
MCYT	0.07	90%	0.5	-	-	-	0.44	25%	0.8
GPDS	0.23	77%	0.6	0.66	56%	0.7	-	-	-

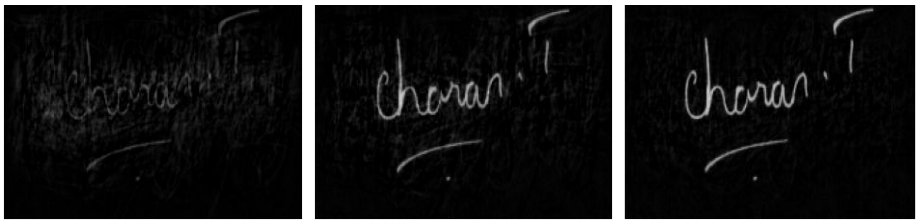

Fig. 5. Effect of Region Restriction on Attack Success Rate and SIIM for the proposed method

6 Conclusion

In this paper, we proposed a black-box transferable attack method to evaluate the robustness of Signature Verification Networks. A novel algorithm to generate a universal noise using ideas from a spatial transformation tool and a lightweight data representation tool is Principal Component Analysis (PCA), followed by restricting the perturbation application area to ensure imperceptibility. It's a complete black-box method with no information about the model architecture or weights learned. The experimental results on three widely used benchmark datasets highlight the strength of the proposed approach. We achieved a high attack success rate of 90%. We also conducted experiments to prove that our

proposed approach is transferable across different datasets as well as model architectures. In the future, we would like to auto-tune the process of region restriction as well as evaluate the proposed attack against various defense systems.

References

1. Alparslan, K., Alparslan, Y., Burlick, M.: Adversarial attacks against neural networks in audio domain: Exploiting principal components. arXiv preprint arXiv:2007.07001 (2020)
2. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248 (2017)
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy. pp. 39–57. IEEE (2017)
4. Dey, S., Dutta, A., Toledo, J.I., Ghosh, S.K., Lladós, J., Pal, U.: SigNet: Convolutional siamese network for writer independent offline signature verification. CoRR **abs/1707.02131** (2017), http://arxiv.org/abs/1707.02131
5. Ferrer, M.A., Diaz-Cabrera, M., Morales, A.: Static signature synthesis: A neuromotor inspired approach for biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 667–680 (2014)
6. Garg, I., Panda, P., Roy, K.: A low effort approach to structured cnn design using pca. IEEE Access **8**, 1347–1360 (2019)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
8. Hafemann, L.G., Sabourin, R., Oliveira, L.S.: Learning features for offline handwritten signature verification using deep convolutional neural networks. Pattern Recognition **70**, 163–176 (2017)
9. Hafemann, L.G., Sabourin, R., Oliveira, L.S.: Characterizing and evaluating adversarial examples for offline handwritten signature verification. IEEE Transactions on Information Forensics and Security **14**(8), 2153–2166 (2019)
10. Jahangir, M., Malik, M.I., Shafait, F.: Adversarial attacks on convolutional siamese signature verification networks. In: International Conference on Document Analysis and Recognition. pp. 350–365. Springer (2023)
11. Kalera, M.K., Srihari, S., Xu, A.: Offline signature verification and identification using distance statistics. International Journal of Pattern Recognition and Artificial Intelligence **18**(07), 1339–1360 (2004)
12. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial Intelligence Safety and Security, pp. 99–112. Chapman and Hall/CRC (2018)
13. Li, H., Li, H., Zhang, H., Yuan, W.: Black-box attack against handwritten signature verification with region-restricted adversarial perturbations. Pattern Recognition **111**, 107689 (2021)
14. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
15. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1765–1773 (2017)
16. Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., Gonzalez, J., Faundez-Zanuy, M., Espinosa, V., Satue, A., Hernaez, I., Igarza, J.J., Vivaracho, C., et al.: Mcyt

baseline corpus: a bimodal biometric database. IEE Proceedings-Vision, Image and Signal Processing **150**(6), 395–401 (2003)
17. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
18. Vorugunti, C.S., Mukherjee, P., Pulabaigari, V., et al.: Osvnet: convolutional siamese network for writer independent online signature verification. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1470–1475. IEEE (2019)
19. Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M., Song, D.: Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612 (2018)
20. Yang, J., Zhang, D., Frangi, A.F., Yang, J.y.: Two-dimensional pca: a new approach to appearance-based face representation and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(1), 131–137 (2004)
21. Zhang, Y., Tian, X., Li, Y., Wang, X., Tao, D.: Principal component adversarial example. IEEE Transactions on Image Processing **29**, 4804–4815 (2020)