SPOOFING ATTACK DETECTION BY ANOMALY DETECTION

Soroush Fatemifar, Shervin Rahimzadeh Arashloo⋆*, Muhammad Awais and, Josef Kittler*

Centre for Vision, Speech and Signal Processing, University of Surrey, UK
⋆Bilkent University, Turkey

ABSTRACT

Spoofing attacks on biometric systems can seriously compromise their practical utility. In this paper we focus on face spoofing detection. The majority of papers on spoofing attack detection formulate the problem as a two or multiclass learning task, attempting to separate normal accesses from samples of different types of spoofing attacks. In this paper we adopt the anomaly detection approach proposed in [1], where the detector is trained on genuine accesses only using one-class classifiers and investigate the merit of subject specific solutions. We show experimentally that subject specific models are superior to the commonly used client independent method. We also demonstrate that the proposed approach is more robust than multiclass formulations to unseen attacks.

Index Terms— Face anti-spoofing, Anomaly detection, client-specific information, one-class classification, Convolutional neural networks

[1]

1. INTRODUCTION

Despite the significant improvements attained in biometric devices, spoofing attacks are recognised as a considerable threat to face recognition systems where an impostor tries to access a service illegally by using a variety of different approaches of face presentation attacks (PA). Although there have been many different types of PAs in spoofing scenarios, print attack, replay attack, and 3D mask are much more accessible to fraudsters and are commonly encountered in practice. Due to diversity of spoofing attacks, there is an increasing need to design a face recognition system that can detect novel and unknown types of spoofing attempts.

In order to counteract PAs, the majority of approaches [2, 3] formulate the problem as two-class classification to learn distinguishable cues separating genuine access and attack attempts. However, the two-class formulation in real-world scenarios does not perform robustly due to its poor generalisation performance in the presence of novel attack types [4]. Another drawback of the two-class classification methods is that finding an optimised boundary to differentiate between various attack categories from genuine access data is not a straightforward task due to the class imbalance problem, the difficulty of collecting spoofing attack samples and cross device variability [1].

To address the aforementioned shortcomings, the authors in [1] propose an anomaly detection system that considers genuine data as normal observations and impostor attacks as anomalous samples. The apparent superiority of anomaly detection learners, so called one-class classifiers, compared to two-class classification methods is that they can be robust to previously unseen and innovative attacks [1]. In this paper we build on this approach, and being inspired by the recent advances of convolutional neural networks (CNN) in the anti-spoofing studies [5, 6, 7], our one-class classifiers are built using the representations obtained from the deep pre-trained CNN models. To examine the capability of different CNN architectures in the anti-spoofing scenarios, face-tuned CNN networks as well as general object classification CNNs are chosen to extract features from video frames.

In addition to one-class SVM and one-class SRC that have been used for spoofing detection previously [1, 4], we also experiment with two more classification approaches namely, Mahalanobis Distance(MD) and Gaussian Mixture Model(GMM) to investigate the capability of probability estimation based models in anti-spoofing scenarios.

Using anomaly approaches and CNN features, we investigate the merits of developing client-specific models as compared to the usual client independent setting, since the identity of a client is known to the biometric system. This mirrors similar attempts in the context of face recognition [8, 9]. The idea of using client-specific information for spoofing detection was first explored in [10]. However, this study related to the classical multiclass formulation of the spoofing detection problem. In this paper, we propose the use of subject specific models in the context of anomaly detection. Moreover, we advocate the use of client-specific thresholds to improve the spoofing detection performance further.

We evaluate the proposed solution on benchmarking spoofing attack datasets, namely Replay-Attack [11] and Replay-Mobile [12] to make a fair comparison with the state of the art. The client-specific variant of the anomaly detection approach requires a new evaluation protocol which we

[1]This work was supported in part by the EPSRC Programme Grant (FACER2VM) EP/N007743/1 and the EPSRC/dstl/MURI project EP/R018456/1.

introduce for experiments on the new and more challenging face spoofing dataset of ROSE-Youtu [3], covering a diverse variety of illumination conditions, camera sources, and attack types.

The main contributions of this paper include: a) Developing anomaly detection solutions for the anti-spoofing task using representations derived by different CNN architectures, b) building client-specific models for spoofing detection, c) adopting client-specific thresholds for each model, d) defining a new evaluation protocol for experimenting on the Rose-Youtu dataset, e) demonstrating experimentally that the proposed client-specific anomaly detection approach with client-specific thresholds delivers superior performance and is more robust to unseen types of attacks.

The rest of the paper is organised as follows: In Section 2, we introduce the proposed class-specific anomaly detection framework for anti-spoofing detection. The experimental results are reported in Section 3. Finally, conclusions are drawn in Section 4.

2. ANOMALY DETECTION

2.1. Anomaly Classifiers

Anomaly detection is the process of finding patterns deviating from the expected behavior defined by "normal" samples of a training dataset. Therefore, the fundamental task of a one-class classifier is to detect anomaly observations among test samples. In the area of face spoofing detection, the anomaly detection problem is often formulated as a two class problem, or occasionally, as a one class decision making problem but with decision thresholds set using some face recognition system attack samples. This paper considers the classical scenario where only genuine access samples are used for training. This "purest" approach is compared with the anomaly detection mechanism in which some attack data is available for the system design to gauge the relative merit of positive class samples on the spoofing attack detection performance.

The anomaly detectors used in this paper include:

One-CLASS SVM: Support Vector Data Description (SVDD) [13] is a widely used one-class extension of SVM. SVDD encloses the normal training data by a minimum radius hypersphere. Outliers are detected as test samples falling outside the hypershpere.

One-CLASS SRC: Sparse representation based classifier [14] is a non-parametric outlier detection method assuming that a test sample can be represented as a sparse linear combination of available training samples. Here, the reconstruction error is used to detect outliers.

One-CLASS MD: Assuming that genuine access data follows a single-mode Gaussian distribution, the Mahalanobis distance of a test sample to the mean can serve as an output of a one-class MD spoofing detector.

One-CLASS GMM: A Gaussian mixture model is a parametric probability density function representing a weighted sum of Gaussian component densities. Its model parameters are estimated using the Expectation Maximization algorithm [15]. Outliers of the model can be flagged by measuring the minimum MD to the respective mixture components.

2.2. client-specific versus Client Independent Anomaly Detection

In the anomaly based spoofing detection approach the one-class classifiers are designed using genuine access data. For each client C_i such *normal* access data X_i can be the enrolment biometric traits, potentially augmented by operational data collected during the live operation of an installed biometric system. Each set X_i contains multiple biometric samples for client C_i. For the conventional, client independent solution, each one-class model is trained using the union X of these client dependent sets, i.e. $X = \cup_i X_i$. In contrast, the client-specific designs are trained using the client-specific sets $X_i, i = 1, ..., n$, where n denotes the number of clients.

A one-class classifier produces a score for each biometric trait. For a set of traits X (X_i) the classifier will generate a set of scores S (S_i). All these scores represent normal accesses. The assumption is that spoofing attacks would generate scores that differ from normal scores, and could be detected as outliers of the distribution of normal scores. For that we need to define a threshold. A common practice is to set the threshold at a predefined level of confidence. This means a threshold that rejects a given small proportion of normal scores, usually 1-15%. If a development set was available, one could set the threshold so that the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are balanced, producing an equal error rate (EER). For a client independent model, the threshold T would be set on the distribution of scores S. In the anomaly formulation adopted here the Half Total Error Rate (HTER) would correspond to the operation point defined by the selected level of confidence. We are reporting the results that correspond to the operation point closest to the EER setting.

For client-specific models we have two options. We can combine all client-specific score sets into a single set $= \cup S_i$ and determine a single global threshold $\hat{T}$ for this merged population of scores. An alternative is to define client-specific threshold T_i for each population of scores S_i.

The major drawback of a single global threshold is that it is only applicable if different clients have similar score distributions which is rarely the case in practice. To demonstrate the significance of a subject specific threshold setting, score distributions of three clients are depicted in Figure 1. As seen in the figure, the subject specific threshold of each client can separate the genuine and attack score distributions, in terms of HTER, quite well, compared to a single global threshold. The superiority of using client-specific thresholds derives from the fact that when data of different clients are combined together,

it becomes more difficult to find a common threshold that satisfies them jointly. The power of subject specific information can be gleaned from HTERs of different clients using global and client-specific thresholds shown in the Table in Figure 1. Accordingly, the worst case scenario for the spoofing problem is the red client having a wide attack distribution and a tight genuine score distribution. It is interesting to note that the biggest difference in HTERs produced by the two types of thresholds is achieved by the red client.

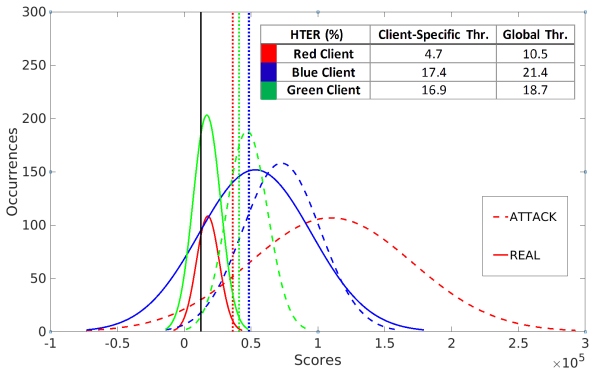

Fig. 1. Score distributions and client-specific thresholds of three different clients using MD classifier and GoogleNet [16] features. The score distributions of genuine and spoofing attack samples of different clients are drawn by solid and dashed lines, respectively. The vertical dashed lines and solid black line represent the client-specific and global thresholds, respectively.

For class-independent approaches, it is apparent that finding an optimal hyper-ellipsoid for SVDD in the presence of diverse score distributions is likely to be more difficult in compared to class-specific methods. In the case of the one-class SRC classifier, one can expect both class-specific and class-independent methods to produce the similar classification model. However, the computational complexity of classifier training is a function of training set size. Accordingly, the convergence time of the class-independent methods is much greater and will be an important consideration in one-class classifier selection. It is expected to see a considerable difference between class-specific and class-independent MDs because the mean of the merged distribution of scores will differ from any individual client-specific score distribution. Regarding the comparison of client-specific and client-independent methods, both approaches have a similar GMM model only if score distributions of different clients are nearly interchangeable, which is not the case in the majority of datasets.

2.3. Access Data Representation

Deep networks such as CNNs have received remarkable attention over the last few years. Inspired by this fact, in the current work, deep pre-trained CNN models are used to extract features from an image. Networks such as GoogleNet [16], ResNet50 [17] are categorised in the generic group while VGG-verydeep-16 [18] and VGG-Face [19] are trained and tuned for a face classification.

3. EXPERIMENTS

3.1. Datasets and Protocols

The experiments are performed on three anti-spoofing datasets namely, Replay-Attack, Replay-Mobile, and Rose-Youtu. The Replay-Attack dataset contains 1,300 genuine and spoofing attack videos from 50 subjects. These videos were captured under two illumination conditions: controlled and adverse. Three different spoofing types are included: print attack, video attack, and digital photo. The Replay-Mobile dataset consists of 1,190 video clips of both real-access and attack attempts of 40 subjects. The sequences were taken using two different acquisition devices and five different mobile scenarios. ROSE-Youtu is the most recent spoofing dataset covering a large variety of illumination conditions, camera models, and attack types. This dataset consists of 3,350 videos for 20 subjects. Five mobile phones and five different illumination conditions were used to collect the dataset.

The enrolment set available for each client is used for training the class-specific approaches. In compliance with the anomaly detection formulation, only real-access samples are used to build both class-specific and class-independent approaches. For the Rose-Youtu dataset, that does not contain the enrolment set for each client, a new evaluation protocol is proposed to implement the class-specific approaches in which 40% of the real-access videos provided for each client is selected to form their enrolment set. The proposed protocol ensures that an enrolment set contains videos of all lighting conditions and camera sources.

3.2. Implementation Details

Before feeding video clips to pre-trained networks, each frame is photometrically normalised based on the retina method [20] to reduce the impact of different lighting conditions. For the Rose-Youtu dataset, face bounding boxes are detected by the Viola-Jones algorithm [21]. To extract representations for the pre-processed video clips, the output of the pre-ultimate layer of each network is used as features. The one-class SVM classifier implementation is based on the SVDD classifier from LIBSVM [22]. The classifier output of the one-class SVM is a spoofing detection score. In the case of the MD, the parameters of the Gaussian distribution of the feature vector are estimated. For each query sample, the Mahalanobis distance is computed as a detection score. For GMM classifier, the minimum Mahalanobis distance of a test sample from all K components is regarded as the spoofing detection score. The number of mixture components K is set for each dataset based on grid search using cross validation.

Table 1. HTER(%) is presented for the test set of Replay-Attack dataset. The best result is marked in bold.

The Replay-Attack Dataset (HTER%)									
	SVM		SRC		MD			GMM	
	Spec	Indp	Spec	Indp	Spec	*Cs-Gb*	Indp	Spec	Indp
GoogleNet	16.17	36.35	16.45	18.15	4.04	*5.92*	17.19	15.89	16.56
ResNet50	11.99	41.66	21.06	19.54	**2.82**	*5.23*	15.59	14.44	15.12
VGG-VD-16	10.24	35.65	16.12	14.65	3.26	*4.8*	13.26	15.03	17.01
VGG-Face	17.33	46.5	19.11	17.45	5.68	*7.27*	12.75	9.96	12.79

Table 2. HTER(%) is presented for the test set of Replay-Mobile dataset. The best result is marked in bold.

The Replay-Mobile Dataset (HTER%)									
	SVM		SRC		MD			GMM	
	Spec	Indp	Spec	Indp	Spec	*Cs-Gb*	Indp	Spec	Indp
GoogleNet	14.34	24.21	21.54	21.78	**13.70**	*15.91*	16.74	14.21	15
ResNet50	21.76	35.69	30.75	32.14	21.81	*22.05*	26.38	21.53	25.77
VGG-VD-16	18.78	30.02	29.65	33.47	19.84	*22.17*	20.52	18.05	21.03
VGG-Face	34.62	35.49	42.54	46.18	33.25	*33.63*	33.23	38.5	32.42

3.3. Results

Different combinations of CNNs, classifiers, and datasets are used to carry out extensive experiments in this section. The performance is measured using HTER. Tables 1 to 3 report the HTER on the test set of Replay-Attack, Replay-Mobile, and Rose-Youtu, respectively. In each Table, "Spec" denotes the client-specific models while "Indp" represents the class-independent approaches. Another column, designated by "Cs-Gb", is added to compare the performance of client-specific models using global and client-specific thresholds. The global threshold results are reported only for the overall best classifier of the Replay-Attack and Replay-Mobile datasets, and for all classifiers of the Rose-Youtu.

As shown in Table 1, class-specific MD using the ResNet50 representation is the best approach, achieving the minimum HTER rate of 2.82%. This drops by at least 1.54% if global thresholds are used instead of the client-specific setting. Compared to the two-class classifier formulation, our lowest HTER of 2.82% is superior to the two-class SVM in [11] by 1.28% and 3.28% in [23]. As is observed in Table 2, class-specific MD with the GoogleNet features achieved the best performance among the other methods with the HTER rate of 13.7%. Again, the performance degrades when using global thresholds. To compare with two-class methods, the proposed anomaly detection system with 13.7% HTER rate is better than 19.87% in [12]. Due to the poor performance on the Replay-Attack and Replay-Mobile datasets, and computational complexity, the SRCs were excluded from experiments on Rose-Youtu. According to Table 3, the lowest HTER of 13.6% is obtained by class-specific GMM using the ResNet50 features. Similar to previous datasets, class-specific thresholds, compared to the global ones, in Rose-Youtu dataset give a better solution. Overall, in one-class SVMs, the gap between the performance of class-specific and class-independent models is huge since the hypersphere optimisation process in the class-independent approaches cannot be accomplished successfully in the presence of diverse feature distributions. For one-class SRCs, the performance of the class-specific models is slightly better than that of class-independent methods due to the fact that each frame can be sufficiently reconstructed by a sparse linear combination of samples. In one-class MDs, class-specific models outperform class-independent ones by a large margin because single mode Gaussian model can better estimate the feature distribution of each individual client. In one-class GMMs, class-specific models are mostly better than class-independent ones due to the fact that the mixture of Gaussian components can cluster the data distribution better when it relates to a single subject. An additional benefit is gained from using subject specific thresholds versus global thresholds.

Table 3. HTER(%) is presented for the test set of Rose-Youtu dataset. The best result is marked in bold.

The Rose-Youtu Dataset (HTER%)									
	SVM			MD			GMM		
	Spec	*Cs-Gb*	Indp	Spec	*Cs-Gb*	Indp	Spec	*Cs-Gb*	Indp
GoogleNet	19.99	*20.38*	25.53	15.8	*17.03*	20.25	15.79	*17.03*	19.87
ResNet50	17.72	*18.08*	23.68	17.23	*18.10*	20.3	**13.6**	*15.87*	18.43
VGG-VD-16	18.73	*18.96*	23.03	16.57	*17.48*	19.51	16.05	*18.95*	20.23
VGG-Face	20.07	*20.81*	31.73	15.92	*17.28*	18.26	17.12	*18.28*	20.49

In summary, class-specific approaches perform consistently better than class-independent methods in experiments. The performance of generic CNN networks is usually better compared to those tuning for the purpose of face recognition. Finally, one-class classifiers outperform conventional multiclass classifiers which confirms the merit of adopting the anomaly detection formulation in spoofing scenarios.

4. CONCLUSIONS.

Motivated by the benefits of the anomaly based approach to face spoofing detection, we proposed a client-specific extension of the method to capitalise on its ability to create sharper genuine access data models. Accordingly, the spoofing attack detectors were designed as one-class classifiers using normal access data exclusively. We also showed that the performance of the detectors can further be enhanced by using client-specific detection thresholds, rather than a global threshold. Among the machine learning tools considered for the one-class classifier design, the Gaussian distribution based statistical hypothesis testing proved most effective, with the exception of one benchmark dataset where its Gaussian mixture model extension proved preferable. Extensive experiments involving three spoofing datasets confirmed the merits of the client-specific anomaly detection approach. The reported results demonstrated that the proposed model yields promising performance compared to the class-independent formulation as well as to conventional multiclass classification models.

5. REFERENCES

[1] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE Access*, vol. 5, pp. 13868–13882, 2017.

[2] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," *CoRR*, vol. abs/1803.11097, 2018.

[3] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 1794–1809, July 2018.

[4] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel, "On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing," in *2018 International Conference on Biometrics (ICB)*, pp. 75–81, Feb 2018.

[5] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *CoRR*, vol. abs/1408.5601, 2014.

[6] M. Sajjad, S. Khan, T. Hussain, K. Muhammad, A. K. Sangaiah, A. Castiglione, C. Esposito, and S. W. Baik, "Cnn-based anti-spoofing two-tier multi-factor authentication system," *Pattern Recognition Letters*, 2018.

[7] L. Li, Z. Xia, L. Li, X. Jiang, X. Feng, and F. Roli, "Face anti-spoofing via hybrid convolutional neural network," in *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, pp. 120–124, Oct 2017.

[8] S. R. Arashloo and J. Kittler, "Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features," *IEEE Transactions on Information Forensics and Security*, vol. 9, pp. 2100–2109, Dec 2014.

[9] S. R. Arashloo, "Multiscale binarised statistical image features for symmetric face matching using multiple descriptor fusion based on class-specific lda," *Pattern Analysis and Applications*, vol. 20, pp. 113–126, Feb 2017.

[10] I. Chingovska and A. R. dos Anjos, "On the use of client identity information for face antispoofing," *IEEE Transactions on Information Forensics and Security*, vol. 10, pp. 787–796, April 2015.

[11] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, pp. 1–7, Sept 2012.

[12] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *Proceedings of the International Conference on Biometrics Special Interests Group (BioSIG)*, Sept. 2016.

[13] D. M. Tax and R. P. Duin, "Support vector data description," *Machine Learning*, vol. 54, pp. 45–66, Jan 2004.

[14] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 210–227, Feb 2009.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, June 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[20] V. Štruc and N. Pavešić, *Photometric normalization techniques for illumination invariance*, pp. 279–300. IGI-Global, 2011.

[21] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, pp. I–I, Dec 2001.

[22] C. chung Chang and C.-J. Lin, "Libsvm: a library for support vector machines," 2001.

[23] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, Dec 2016.