A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira and A. Uhl (Eds.): BIOSIG 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021

Learning by Environment Clusters for Face Presentation Attack Detection

Tomoaki Matsunami[1], Hidetsugu Uchida[2], Narishige Abe[3], Shigefumi Yamada[4]

Abstract: Face recognition has been used widely for personal authentication. However, there is a problem that it is vulnerable to a presentation attack in which a counterfeit such as a photo is presented to a camera to impersonate another person. Although various presentation attack detection methods have been proposed, these methods have not been able to sufficiently cope with the diversity of the heterogeneous environments including presentation attack instruments (PAIs) and lighting conditions. In this paper, we propose Learning by Environment Clusters (LEC) which divides training data into some clusters of similar photographic environments and trains bona-fide and attack classification models for each cluster. Experimental results using Replay-Attack, OULU-NPU, and CelebA-Spoof show the EER of the conventional method which trains one classification model from all data was 20.0%, but LEC can achieve 13.8% EER when using binarized statistical image features (BSIFs) and support vector machine used as the classification method.

Keywords: face anti-spoofing, presentation attack detection, face image clustering.

1 Introduction

Face recognition has been used widely such as access control of personal use devices and the border controls because of its high accuracy and convenience. However, it is vulnerable to Attack Presentations (APs) that try to impersonate someone by presenting a photo, a video, or other item. In recent years, since target face images such as photos to impersonate can be easily obtained through social networking service, attackers can easily try to authenticate using the obtained photos as PA. That is why the presentation attack detection (PAD) could be mandatory to realize the secure face recognition.

The PAD approaches include hardware-based or software-based, and the software-based approach can be divided into motion-based method and methods using image features [RB17]. The hardware-based approach uses dedicated equipment to obtain the information for classification real and fake, Raghavendra et al. used light field cameras [RRB15] and Kose et al. used depth cameras [KD13]. As a motion-based method, Kollreider et al. proposed a method using facial expression changes [KFF07] and

[1] Fujitsu Limitd, 4-1-1 Kamikodanaka Nakahara-ku Kawasaki Kanagawa Japan, t.matsunami@fujitsu.com
[2] Fujitsu Limitd, 4-1-1 Kamikodanaka Nakahara-ku Kawasaki Kanagawa Japan, u.hidetsugu@fujitsu.com
[3] Fujitsu Limitd, 4-1-1 Kamikodanaka Nakahara-ku Kawasaki Kanagawa Japan, abe.narishige@fujitsu.com
[4] Fujitsu Limitd, 4-1-1 Kamikodanaka Nakahara-ku Kawasaki Kanagawa Japan, yamada.shige@fujitsu.com

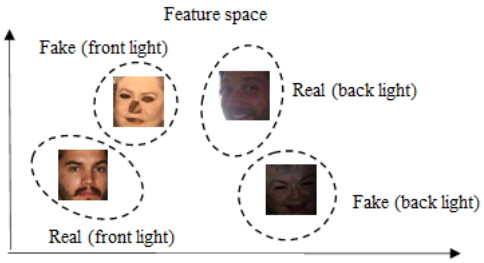

Fig. 1. Illustration of face feature distribution acquired with different light sources

Frischholz et al. proposed a method using head-pose [FW03]. As a method for realizing PAD against a RGB camera image commonly used in face recognition, there are many methods using image features [Wa13]. At first, there are many methods to classify images into two classes of real and fake based on texture features such as local binary feature (LBP) by Pereira et al. [Pe 13]. Haralick feature (HF) by Agarwal et al. [ASV 16], and binarized statistical image features (BSIF) by Soler et al. [SBB 20] and Boulkenafet et al. evaluated combinations of multiple texture features [BKH 18]. Moreover, many techniques using deep learning have also been proposed, such as convolutional neural network (CNN) based [YLL14, PHJ16] and long short-term memory (LSTM) [SSL18]. On the other hand, since various PAIs are used, there is an issue that fake images have a large diversity, and the distribution of fake features is complicated. To solve this issue, 1-class classification method which trains only 1 class of real has been proposed. In the 1-class classification, by training fake as an outlier, real can be identified with high accuracy even if fake feature distribution becomes complicated. Agarwal et al. proposed 1-class SVM [AKW 17], Nikisins et al. proposed 1-class GMM [Ni 18], Arashloo proposed 1-class FV [AR 20], and Bawja et al. proposed 1-class CNN method [Ba 20].

However, in the actual face recognition, there are factors that increase the distribution of features for both real and fake. One of the most prominent examples is the change in the lighting environment, such as the position and intensity of the light source, and this paper focuses on the change of light. For example, as illustrated in Fig. 1, the distance due to the change in the lighting environment can be larger than the distance between real and fake in the feature space. In the existing methods using image features in such a case, the fake is included in the real distribution and the real is included in the fake distribution, and the boundary between the real and the fake cannot be accurately estimated. Therefore, the purpose of our study is to focus on PAD using image features from RGB images, and to train a model that can accurately classify between real and fake, even for training data that includes variations of PAIs and lighting environment. Hereinafter, in this paper, PAI and lighting environment are referred to as photographing environment. The main contributions of our work are as follows:

- We propose Learning by Environment Clusters (LEC) which divides training data into some clusters of similar photographic environments and trains real and fake classification models for each cluster.

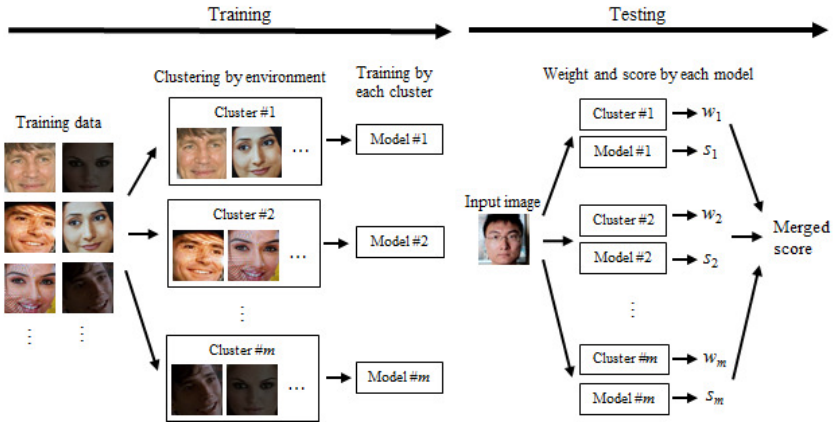

Fig. 2. Process flow of the proposed method

- Using hierarchical cluster analysis, LEC can classify training data into environment clusters even when the photographing environment of them is not known.
- We confirmed the effectiveness of the proposed method on HF+SVM, BSIF+SVM, and BSIF+CNN as a classification method using three PAD datasets, Replay-Attack [CAM12], OULU-NPU [Bo17], and CelebA-Spoof [Yu20].

The structure of this paper is as follows. Chapter 2 shows the protocol of the proposed method. Chapter 3 describes the evaluation results of the proposed method. Finally, Chapter 4 summarizes the results and discusses future prospects.

2 Learning by Environment Clusters

Fig. 2 shows the process flow of the proposed method. In the training phase, the training data is divided into clusters for each similar photographing environment, and a classification model is trained for each cluster. It is assumed that the training data includes images acquired under various photographing environments and that the photographing environment of each image is not given. When the PAI or the lighting environment changes, the way of reflection changes and appears as a local change. LBP can represent the local change in the image, that's why we adopt the LBP as a descriptor. Since the distance of the LBP histogram becomes small between images acquired in the similar environment, clustering is enabled even when the photographing environment of the training data is not given by using the LBP histogram. In the LBP histogram, the $LBP_{8,1}$ proposed in [OPM02] is calculated from the HSV and YCbCr channels of the face image, and a 354-dimensional feature vector is obtained. Next, the training images are divided into clusters using LBP histograms by hierarchical cluster analysis (HCA) [Wa63]. In HCA, combining the most similar clusters is repeated with N clusters each including one LBP histogram as a start. The Ward's method was used to calculate the

Algorithm Algorithm for hierarchical cluster analysis

Input: N LBP histograms
Output: Clusters by environment
Initialize: Generate N clusters, each of which contains 1 LBP histogram and set nc to the number of clusters ($nc = N$)
Definition: cluster distance $cd(C_i, C_j) = L(C_i \cup C_j) - L(C_i) - L(C_j)$, where $L(C_i)$ is sum of squares of Euclid distance between each LBP histogram contained in C_i and LBP histogram centroid of C_i
1 Calculate mcd as minimum of cd for all combinations of 2 clusters
2 **while** $mcd <$ threshold of cd
3 Conbine C_i and Cj where $cd(C_i, C_j)$ is equal to mcd into C_k, where k is new ID
4 Update $L(C_k) \leftarrow L(C_i \cup C_j)$ and $nc \leftarrow nc - 1$
5 Update mcd in new nc clusters
6 **end** while

similarity between two clusters. The algorithm for HCA is shown in Algorithm 1. HCA can divide training data into the arbitrary number of clusters by reaching at the end condition. In this paper, the termination condition was determined as the time when the minimum value of cd exceeded the threshold value from the pre-experimental results. By using HCA, even when it is not known how many face images of what kind of photographing environment are included in the training data, a cluster for each similar photographing environment can be generated. In the following, the number of clusters in the generated photographing environment is represented by m, and the i-th cluster is represented by C_i. It then learns a model that classifies real and fake for each of the m clusters. There is no particular limitation on the learning method of the model. In this paper, we verified SVM learning using hand-crafted features and CNN learning. As a result of the learning phase of the proposed method, a model specific to the photographing environment indicated by each cluster is generated, and the model learned by the cluster C_i is represented by M_i. In the test phase, a score S_i, which is an output of the model M_i for an input image, and a weight w_i of the score are calculated in an i ($1 \leq i \leq m$) th photographing environment clustered during learning. w_i is calculated as $w_i = 1/Ed_i$, where Ed_i is Euclidean distance between the LBP histogram of the input image and the center of gravity of C_i. By making the w_i larger when the distance between the input image and C_i is small, even if it is unclear in what kind of shooting environment the input image was acquired, real and fake can be determined for various photographic environments by calculating a merged score where the weight of the result of a model in a similar photographic environment is increased from the training data.

3 Experiments

The experimental protocol aims to address that our proposed method (LEC) is effective regardless of datasets or classification models. The experimental evaluation

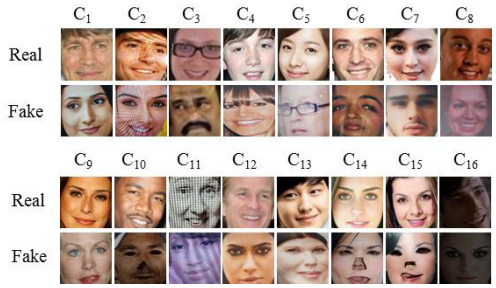

Fig. 3. Image examples in the each cluster

was conducted over three PAD datasets, Replay-Attack [CAM12], OULU-NPU [Bo17], and CelebA-Spoof [Yu20]. Replay-Attack includes photos and videos (replay), OULU-NPU includes prints and videos, and CelebA-Spoof includes prints, videos, and paper cuts. For each data set, MTCNN [Zh16] is used as a face detector, and each detected face is aligned to a size of 112 x 112. Replay-Attack and OULU-NPU are video data sets, so we extract one frame of face image per a video. Either OULU-NPU or CelebA-Spoof is used for training, and CelebA-Spoof is used for testing. When the minimum value of *cd* is more than 200 in HCA, the clustering is terminated to generate four clusters for Oulu-NPU and 16 clusters for CelebA-Spoof. Fig. 3 shows examples of face images included in each cluster C_i ($1 \leq i \leq 16$) of the CelebA-Spoof. The photographic environment is divided into clusters such that C_1 is less affected by the lighting, C_{11} is shaded on the surface, and C_{16} is backlit. In terms of the classification model, we use three models generated by HF + SVM, BSIF + SVM, and BSIF + CNN. As described in [ASV16] for HF, RDWT is applied to each RGB channel of a face image to obtain four sub-bands, and the original image and the four sub-bands are divided into 3x4 patches respectively, and 13 features are calculated from each patch to obtain feature vectors with 2,340 (= 3 x 5 x (3x4) x 13) dimensions. For BSIF, there are 60 pre-learned filters from natural images [KR 12]. We use one of the filters with a filter size of 7 x 7 and a filter number of 8 and obtain 768-dimensional feature vectors by extracting features from 6 channels of HSV and YCbCr. The SVM determined the hyperparameters by 5-fold cross validation using a linear kernel. CNN is a one-dimensional CNN having three convolution layers with 768-dimensional feature vectors of BSIF as input, and a batch normalization is inserted immediately after a second or third convolution layer. The activation function is Leaky ReLU (Negative slope factor is 0.2) and the dropout rate was 0.25. First, a model for each cluster is obtained by transfer learning of all coupling layers for the data of each cluster using a model trained with all data.

We use the evaluation protocol based on the international standard ISO/IEC 30107-3 [ISO17] using attack presentation classification error rate (APCER) which indicates the rate at which the attack (fake) is erroneously determined as Bona-Fide (real), the bona fide presentation classification error rate (BPCER), which indicates the rate at which the bona fide is erroneously determined as fake, the equal error rate (EER), which is the error rate at APECR = BPCER, and the half total error rate (HTER), which is the average of APECR and BPCER.

Tab.1 Evaluation results on LEC using public datasets

Train	Test	Classification method	EER		HTER	
			woLEC	wLEC	woLEC	wLEC
OULU-NPU	Replay-Attack	HF + SVM	45.0%	41.5%	44.6%	39.6%
CelebA-Spoof	Replay-Attack	HF + SVM	23.8%	22.5%	22.4%	20.6%
OULU-NPU	Replay-Attack	BSIF + SVM	46.3%	37.8%	46.0%	36.6%
CelebA-Spoof	Replay-Attack	BSIF + SVM	20.0%	13.8%	16.8%	13.8%
CelebA-Spoof	Replay-Attack	BSIF + CNN	20.0%	16.3%	19.1%	15.9%

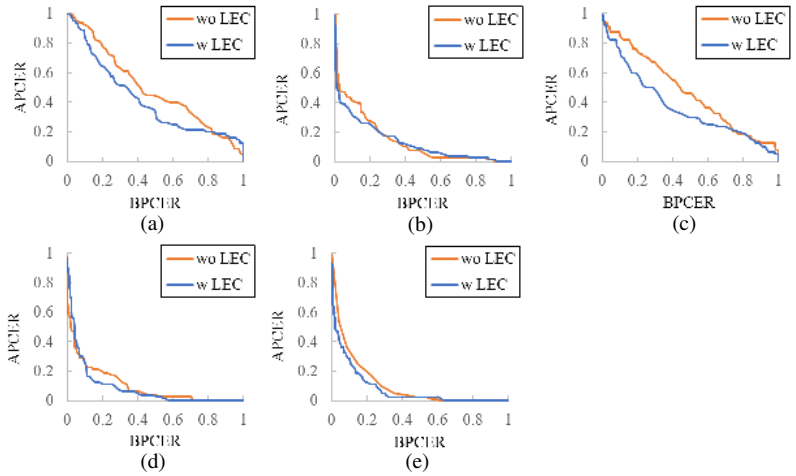

Fig. 4. DET curves of (a) HF + SVM trained with Oulu-NPU, (b) HF + SVM trained with CelebA-Spoof, (c) BSIF + SVM trained with Oulu-NPU, (d) BSIF + SVM trained with CelebA-Spoof, and (e) BSIF + CNN trained with CelebA-Spoof.

The evaluation results are shown in Tab. 1. by comparing proposed method (wLEC) with conventional method when one classification model is generated using all the training data (woLEC). Fig. 4 shows the respective detection error tradeoff (DET) curves. LEC improves both EER and HTER for both datasets and feature extraction methods. From these results, it can be said that the classification model generated from each cluster by LEC is specialized for each photographic environment, and that by increasing the weight of the result obtained from the model of the photographic environment similar to the input image, a robust classification method for the change of the photographic environment can be realized. Further, in the proposed technique, the optimal number of divisions of the training data depends on the diversity of the photographic environment of the training data, and when the number of divisions is small with respect to the diversity of the training data, the model is learned from a data set having a large diversity, and a problem similar to the case of no division occurs. On the contrary, when the number of divisions is large, a plurality of models of similar environments are learned, and the effect of division is reduced. Therefore, it is considered optimal to dynamically determine the number of partitions according to the distance between clusters as we proposed. LEC can be applied regardless of the feature

extraction method or discriminator, therefore we used typical hand-crafted features, SVM, and shallow CNN for evaluation in this paper. Experimental results show the effectiveness of LEC in all combinations, hence we consider that LEC is effective even when the state-of-the-art method is used for feature extraction and discriminator.

4 Conclusion

In this paper, in order to deal with the various Presentation Attack Instruments and various photographing environment by lighting environment at the time of face image acquisition in PAD, we proposed LEC which divides training data into clusters depending on the photographing environment and trains the classification model of real and fake for each cluster. Experimental results using Replay-Attack, OULU-NPU and CelebA-Spoof shows the effectiveness of the proposed method, for example, EER of the conventional method, which trains one classification model from all data, was 20.0%, while the accuracy was improved to 13.8% EER using LEC in the case of using BSIF + SVM. As a future work, the proposed method is evaluated by focusing on photographs and videos using the 2-class classification method in this paper, however, it is necessary to confirm whether the proposed method can be used in a more general way by verifying it using the 1-class classification method and evaluating it using various datasets of different types of presentation attack such as 3D face masks.

References

[AKW17] Arashloo, S. R.; Kittler, J.; Christmas, W.: An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. IEEE Access, 5: pp. 13868 -13882, 2017.

[Ar20] Arashloo, S. R.: Unseen face presentation attack detection using sparse multiple kernel fisher null-space, IEEE Trans. on Circuits and Systems for Video Technology, 2020.

[ASV16] Agarwal, A.; Singh, R.; Vatsa, M.: Face Anti-Spoofing using Haralick Features, IEEE 8th Int. Conf. on biometrics theory, applications and systems, 2016.

[Ba20] Bawja, Y. et. al.: Anomaly Detection-Based Unknown Face Presentation Attack Detection, International Joint Conference on Biometrics 2020.

[BKH18] Boulkenafet, Z.; Komulainen, J.; Hadid, A: On the generalization of color texture-based face anti-spoofing. Image and Vision Computing, 2018

[Bo17] Boulkenafet, Z. et. al.: OULU _ NPU: A mobile face presentation attack database with real-world variations, in Proc. FG, pp. 612 – 618, 2017.

[CAM12] Chingovska, I.; Anjos, A.; and Marcel, S.: On the efficiency of local binary patterns in face anti-spoofing, International Conference of the Biometrics Special Interest Group 2012.

[FW03] Frischholz, R. W.; Werner, A.: Avoiding replay-attacks in a face recognition system

using head-pose estimation, IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003.

[ISO17] ISO/IEC JTC 1 SC 37 Biometrics. ISO/IEC FDIS 30107 -3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting. The International Organization for Standardization, 2017.

[KD13] Kose, N.; Dugelay, J.-L.: Countermeasure for the protection of face recognition systems against mask attacks, IEEE Intl. Conf. on Automatic Face and Gesture Recognition, 2013.

[KFF07] Kollreider, K: et. al.: Real-time face detection and motion analysis with application in liveness assessment, IEEE Trans. Inf. Forensics Security. 2, 2007.

[KR12] Kannala, K.; Rahtu, E.: BSIF: Binarized statistical image features, in Proc. IEEE International Conference on Pattern Recognition, Nov. 2012, pp. 1363 – 1366.

[Ni18] Nikisins, O. et. al.: On efficiency of anomaly detection approaches against unseen presentation attacks in face anti-spoofing, International Conference on Biometrics, 2018

[OPM02] Ojala, T.; Pietikäinen, M.; Mäenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence. 24 (2002) pp. 971 – 987.

[Pe13] Pereira, T. de Freitas, et. al.: Can face anti-spoofing countermeasures work in a real world scenario?, International Conference on Biometrics, 2013.

[PHJ16] Patel, K.; Han, H.; Jain, A. K.; Cross-database face anti spoofing with robust feature representation, Chinese Conference on Biometric Recognition, 2016.

[RB17] Raghavendra, R.; Busch, C.: Presentation Attack Detection Methods for Face Recognition Systems: A Comprehensive Survey. ACM Computing Surveys, 2017.

[RRB15] Raghavendra, R.; Raja, K. B.; Busch, C.: Presentation attack detection for face recognition using light field camera, IEEE Trans. Image Process. 24, 2015.

[SBB20] Soler, L. J. G.-; Barrero, M. G.-; Busch, C.: Fisher Vector Encoding of Dense-BSIF Features for Unknown Face Presentation Attack Detection, International Conference of the Biometrics Special Interest Group, 2020.

[SSL18] Sun, Z.; Sun, L.; and Li, Q.: Investigation in Spatial-Temporal Domain for Face Spoof Detection, International Conference on Acoustics, Speech and Signal Processing 2018.

[Wa63] Ward, J. H.: Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association. 58 (301), pp. 236–244, 1963.

[Wa13] Waris, M.A. et. al.: Analysis of textual features for face biometric anti-spoofing, the 21st European Signal Processing Conference, 2013.

[YLL14] Yang, J.; Lei, Z.; and Li.S.Z.: Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv: 1408.5601, 2014.

[Yu20] Yuanhan, Z. et. al.: CelebA-Spoof: Large-Scale Face Anti-Spoofing Dataset with Rich Annotations. European Conference on Computer Vision 2020.

[Zh16] Zhang K. et. al.: Joint face detection and alignment using multitask cascaded convolutional networks. Signal Processing Letters (SPL), pp. 1499 – 1503, 2016.