On Effectiveness of Anomaly Detection Approaches against Unseen Presentation Attacks in Face Anti-Spoofing

Olegs Nikisins, Amir Mohammadi, André Anjos, Sébastien Marcel
Idiap Research Institute
Rue Marconi 19, CH - 1920, Martigny, Switzerland
{olegs.nikisins, amir.mohammadi, andre.anjos, sebastien.marcel}@idiap.ch

Abstract

While face recognition systems got a significant boost in terms of recognition performance in recent years, they are known to be vulnerable to presentation attacks. Up to date, most of the research in the field of face anti-spoofing or presentation attack detection was considered as a two-class classification task: features of bona-fide samples versus features coming from spoofing attempts. The main focus has been on boosting the anti-spoofing performance for databases with identical types of attacks across both training and evaluation subsets. However, in realistic applications the types of attacks are likely to be unknown, potentially occupying a broad space in the feature domain. Therefore, a failure to generalize on unseen types of attacks is one of the main potential challenges in existing anti-spoofing approaches. First, to demonstrate the generalization issues of two-class anti-spoofing systems we establish new evaluation protocols for existing publicly available databases. Second, to unite the data collection efforts of various institutions we introduce a challenging Aggregated database composed of 3 publicly available datasets: Replay-Attack, Replay-Mobile and MSU MFSD, reporting the performance on it. Third, considering existing limitations we propose a number of systems approaching a task of presentation attack detection as an anomaly detection, or a one-class classification problem, using only bona-fide features in the training stage. Using less training data, hence requiring less effort in the data collection, the introduced approach demonstrates a better generalization properties against previously unseen types of attacks on the proposed Aggregated database.

1. Introduction

Thankfully to the progress made in the development of advanced learning paradigms, such as Deep-learning, the recognition performance of facial biometric systems has recently improved significantly even in fully unconstrained environments [15, 17]. However presentation attacks (PA), also known as spoofing attacks, pose additional challenges on the path of wide deployment of biometric systems. According to recent research on vulnerability analysis of face recognition systems [13, 16] the facial biometrics is not an exclusion. In the scenario of face PA, an impostor tries to present artificial sample representing biometric characteristics of the face, with an intention to affect the normal operation of the biometric system. Typical face PAs are print and replay attacks. However they are not limited to that, attackers are getting more creative as new technologies appear.

Up to date, the most of research in the field of face presentation attack detection (PAD), was considering the task as a two-class classification problem. In this scenario, samples of both *bona-fide class* (also called live or real samples) and the *class of PAs* are intensively collected, and the two-class classifier is learned to predict the class of the input samples. There is a lot of research done around this fundamental idea, with some notable ideas introduced in [6, 11, 12, 18]. The main focus has been on boosting the anti-spoofing performance for databases with identical types of attacks across both testing and evaluation subsets. Despite many successes achieved up to now, this approach has drawbacks. First, in realistic applications, spoofing attacks may have a very diverse nature and most probably are not present in the training stage. This diversity may be caused by various reasons, for example, different replay devices, environmental changes or novel types of attacks. As a result, a space occupied by feature vectors coming from PA class can potentially be broad. Thus, the decision boundary of two-class classifier learned in the training stage might fail to generalize in operation mode when unseen types of attacks are presented.

The unpredictable nature of attacks poses a need to study, and address if necessary, the generalization properties of PAD systems under unseen types of attacks. In fact, some researchers already highlighted, that face PAD systems using the two-class classification idea, are failing to generalize across both different datasets and unseen PAs. Authors in [8] trained an LBP and SVM based face PAD systems on the CASIA FASD [19], and tested on the Replay-Attack database [6], and vice-versa. The increase of error rates by at least 100% in all cases, has experimentally demonstrated a poor generalization performance of the system. A broad research on the topic was recently introduced in [4] and [3], where authors test the robustness of 20 different PAD algorithms against unseen types of attacks in both intra and inter database experiments. Nearly all possible combinations of types of attacks and datasets are tested within 3 selected databases - Replay-Attack [6], MSU MFSD [18] and CASIA FASD [19]. While one-class classifiers have been used in the past in fingerprint [10] and speaker [1] spoof detection, to the best of our knowledge, [4] is the only paper testing the applicability of one-class classifiers in the task of face PAD. Authors clearly conclude, that, first, anomaly detection based systems are not inferior compared with two-class analogs. Second, neither one-class or two-class face PAD systems perform well enough requiring more research in this direction. However, the paper has some drawbacks, first, the best performing anomaly detection based setup in [4] is based on one-class SVM. In our, more challenging experiments, we demonstrate relatively weak performance of one-class SVM based algorithm, moreover, our proposed GMM-base anomaly detector outperforms one-class SVM by a large margin. Second disadvantage of [4] is the evaluation methodology, all results are reported as an area under the ROC curve without providing the ROC themselves. In our proposed evaluation protocols, we stick to the methodology widely accepted by the research community, reporting both $HTER$ making the results comparable with legacy systems, and DET curves introduced in ISO/IEC $30107-3$ standard.

Motivated by the discussions disclosed above, the current work, first, demonstrates that nearly state-of-the-art face PAD system using two-class classifier is failing to generalize against unseen types of attacks. Second, we propose an anomaly detection, or one-class classifier, based face PAD system having a better generalization properties against unseen types of PAs.

In order to support above statements, the following **main contributions** are proposed in this work. *First*, to demonstrate the generalization issues of two-class anti-spoofing systems we introduce a new challenging *Aggregated database*, which is composed of three publicly available face PAD databases, as well as develop appropriate evaluation protocols for it. The protocols guarantee, that types of attacks in the training and development sets are not present in the evaluation set. Thus, the reported performance is reflecting the behavior of the PAD system under unseen attacks. The proposed Aggregated database unites the data collection efforts of various institutions being a composition of three publicly available datasets: Replay-Attack [6], Replay-Mobile [7] and MSU MFSD [18]. As will be demonstrated in the experimental section, the Aggregated database is more challenging than any of its individual components. Having a wider spectrum of both photo and replay attacks, the database makes a step towards more realistic evaluation scenario.

Second, we propose a face PAD system approaching presentation attack detection as anomaly detection, or a one-class classification problem, using only bona-fide features in the training stage. The system is built of Image Quality Measures (IQM), from [12] and [18], forming a feature space, and a Gaussian Mixture Model (GMM) trained to represent the probability distribution of *bona-fide* samples. In the prediction step, both real and attack samples are classified using this pre-trained GMM. Experiments on the Aggregated database demonstrate a better generalization properties of the introduced PAD algorithm against unseen types of attacks, as opposed to observed PA detectors based on two-class classifiers. Researchers in [4] demonstrate the comparable generalization performance against unseen types of attacks for PAD systems based on one-class and two-class SVMs. Our proposed system outperforms the one-class SVM based by a large margin, which is examined in the experimental section. Additional motivating point of anomaly detectors is the reduced amount of data required for training, hence consuming less effort in the data collection stage. *Finally*, the results reported in this work are fully reproducible: the publicly available databases are used in experiments, the evaluation protocols are strictly defined, and the source code for replicating experiments is published[1].

2. Proposed anomaly detection based face PAD approach

This section briefly introduces the proposed anomaly detection based face PAD algorithm. The *main focus* of the proposed design is to have better generalization properties against unseen types of attacks as opposed to the analogous builds with two-class classifiers, which is proved experimentally in Section 3. The system is composed of three main blocks: a preprocessor, a feature extractor, and a one-class classifier. The preprocessor is cropping and normalizing the faces in the input frames. The feature extractor uses the IQM features introduced in [12] and [18]. The classifier is a GMM based anomaly detector classifying the features into bona-fide or attack classes. IQMs of [12] and [18] has been selected as discriminative features in the proposed PAD system being the best performing among observed baseline algorithms. However, having successful, domain-specific features is not sufficient for good gener-

[1] Code: https://pypi.python.org/pypi/bob.pad.face

alization properties. These features were originally introduced in combination with two-class classifiers to perform detection of PAs. As demonstrated in the next section, original setup doesn't generalize well against unseen types of attacks. We show, that anomaly detectors, or one-class classifiers, suit the needs better.

According to the taxonomy in [14] anomaly detectors can be split into two groups: generative and non-generative. Generative methods tend to model the distribution functions of observations. The non-generative approaches are learning the decision boundary grouping the normal observations.

Authors in [4] demonstrate the comparable effectiveness of non-generative and two-class based method, specifically one-class and two-class SVM, in the inter-database and unseen attack tests. In our experiments, employing a more challenging Aggregated database as opposed to [4], one-class SVM doesn't perform well. On the other hand, our proposed setup using one-class GMM, associated to generative methods, generalizes significantly better against unseen types of attacks.

A Gaussian Mixture Model is a weighted sum of K multivariate Gaussian distributions:

$$p(x|\Theta) = \sum_{k=1}^{K} w_k \mathcal{N}(x; \mu_k, \Sigma_k), \quad (1)$$

where $\Theta = \{w_k, \mu_k, \sigma_k\}_{\{k=1,\dots,K\}}$ are the weights, means and the covariances parameterizing the GMM model. The GMM is learned using the Expectation Maximization (EM) algorithm introduced in [9]. In the training it is assumed, that each component in (1) has its own general covariance matrix. The EM algorithm is a local optimization method. Thus, the quality of the solution depends on the quality of the initial values of the parameters. In our case, the initialization method is based on k-means to pre-cluster the samples. It sets the means and covariances of the Gaussian distributions to the values of means and covariances of each k-means cluster, and also sets the prior probabilities to be proportional to the mass of each cluster. In our case, GMM is trained using IQM feature vectors of the *bona-fide* class only. Once the training is completed, the final score of a sample x, which in our case is the IQM vector, is computed as follows:

$$score = log(p(x|\Theta)) \quad (2)$$

3. Experiments

This section covers details on the proposed Aggregated database and evaluation protocols, following by experimental results for legacy face PAD systems, and the results for the proposed anomaly detection based face PAD setup.

Number of clients	125
Number of videos	2510
Bona-fide videos	660
PA videos	1850
Video resolution	240 × 320, 720 × 1280, 640 × 480, 720 × 480
Print attacks	A4 and A3 prints
Replay attacks	PC matte-screen, iPad Air, iPad 1st., iPhone 3GS, iPhone 5S

Protocols: number of real and attack (att.) videos						
Set	training		develop.		evaluation	
Num. of clients	37		41		47	
Video type	real	att.	real	att.	real	att.
grandtest	200	552	240	616	220	682
photo-photo-video	200	344	240	392	220	268
video-video-photo	200	208	240	224	220	414

Table 1. The main statistics of the proposed Aggregated database.

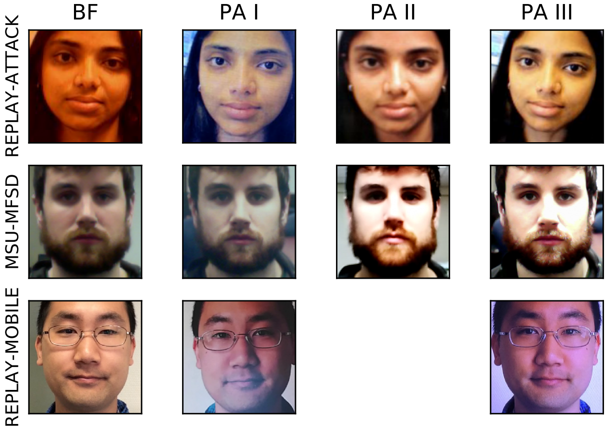

Figure 1. Examples of attacks for databases used in the Aggregated dataset. First column corresponds to bona-fide samples, PA I refers to printed photo attacks, PA II is low quality video replay attacks, and PA III corresponds to high quality video replay attacks.

3.1. Database and evaluation methodology

Subsection covers the introduced Aggregated database and evaluation protocols, allowing to test face PAD systems under unseen attack scenarios. Aggregated database is a composition of three publicly available databases: Replay-Attack [6], Replay-Mobile [7] and MSU MFSD [18], some examples of bona-fide and attack images are displayed in Figure 1. The training, development and evaluation subsets of three forming databases are concatenated producing same subsets of the Aggregated database. Note, the clients in these subsets don't overlap. The main statistics of the Aggregated database is summarized in Table 1. Wide variety of clients, capturing devices and types of attacks makes this database among the largest and most challenging datasets freely available for face PAD experiments. Moreover, using

$HTER$, %	Replay-Attack	Replay-Mobile	MSU MFSD	Aggregated database
LBP [6]	15.6	17.2	21.4	19.1
Motion [2]	13.2	10.4	17.1	43.0
IQM [12, 18]	**4.6**	**4.1**	**4.9**	**15.3**

Table 2. $HTER$ for 3 types of features using two-class SVM with RBF kernel; computed on the evaluation subset of 4 databases for the *grandtest* protocol.

previously published and popular databases as aggregation components as well as similar evaluation protocols, helps researchers to intuitively position this database in the hierarchy. With the database we introduce three evaluation protocols, allowing evaluation of face PAD systems in the scenario of unseen types of attacks. In particular the following protocols have been developed:

- **grandtest** - following the legacy evaluation strategies, in this protocol samples of *all types of attacks* are available *in all subsets* of the database: training, development and evaluation.
- **photo-photo-video** - in this protocol *only photo* attacks are available in the training and development sets, and *only video* attacks are available in the evaluation set.
- **video-video-photo** - the opposite of the above: *only video* attacks are available in the training and development sets, and *only photo* attacks are available in the evaluation set.

It is worth mentioning, that the category of *photo PAs* includes both printed facial photos and photos replayed with screens. Training set is used for training the PAD system. The threshold corresponding to the desired operation point is selected on the development set. The evaluation set is used to report the performance corresponding to the determined threshold. Thus, in protocols *photo-photo-video* and *video-video-photo*, the final performance is reported for types of PAs unseen in the training and development stages.

For all protocols anomaly detectors are trained using purely feature vectors of the bona-fide, or real class. The two-class systems are trained with the data whichever available in the training set of the particular protocol. The software for database querying is publicly available.

3.2. Results for baselines

In this subsection, some successful, previously published face PAD systems are evaluated using three publicly available databases: Replay-Attack, Replay-Mobile, MSU MFSD, as well as using proposed Aggregated database. In this evaluation scenario all types of attacks are present in all subsets of the databases, which corresponds to the *grandtest* protocol available in all aforementioned databases. The goal of these experiments is two-fold. First, the best performing features will be selected for further tests in the unseen attacks scenario, Section 3.3. Second, testing the legacy systems on the Aggregated database helps to intuitively interpret the complexity of the proposed database.

The following systems are selected: LBP-based from [6], Motion-based [2], and IQM-based where feature vector of a frame is a concatenation of quality measures introduced in [12] and [18]. For the sake of compatibility, the preprocessor and classifier are identical in all systems. Since biometric samples in all databases are videos, the preprocessor extracts faces in all frames given annotations defining facial region. The frames with a face smaller than 50×50 pixels are discarded. The cropped images are the then normalized to the identical size of 64×64 pixels. In the case of LBP and Motion features, the facial images are converted to gray-scale format, while IQM-based system requires RGB images. In the LBP based system the feature vectors are normalized LBP histograms composed of uniform $LBP_{8,1}^{u2}$ codes, identical to [6]. In the Motion-based PAD [2], the feature vectors of the length 10 are computed for each 20 *non-overlapping* frames. The IQM system uses 139 quality measures, which is a concatenation of IQMs introduced in [12] and [18]. In all cases the features are classified using identical *two-class* SVM module with RBF kernel. The score is the probability of a sample being a real class. The evaluation of LBP and IQM based systems is done on the *frame-level*, meaning that each frame in the input video is a considered as a separate biometric sample. The Motion-based PAD is evaluated on the *window-level*, each window (20 frmaes) is an individual biometric sample.

The results for this sequence of experiments are introduced in Table 2, reporting $HTER$ on the *evaluation* sets of the databases. The $HTER$ is an average of $BPCER$ and $APCER$ for the threshold, corresponding to EER operation point on the *development* set in this work. The notations $APCER$ - Attack Presentation Classification Error Rate and $BPCER$ - Bona-fide Presentation Classification Error Rate are introduced in ISO/IEC $30107-3$ standard and are similar to FAR and FRR, respectively.

Under identical preprocessor and classifier instances, IQM clearly outperform other types of features. Moreover, one can notice, that the proposed Aggregated database is the most challenging one.

3.3. Results for the proposed PAD system

Following the findings from previous experiment, the generalization properties of one-class and two-class PAD systems under *unseen types of attacks* are studied here, using best performing IQM features and most challenging Aggregated database as an evaluation basis. The following *one-class* classifiers are tested: one-class GMM, which cor-

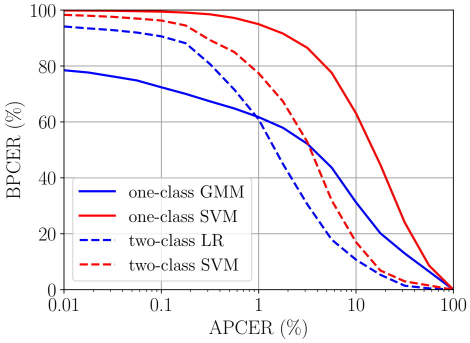

Figure 2. DET curves - *development* set of the **grandtest** protocol of the Aggregated database. Four types of classifiers applied to IQM features of the facial region.

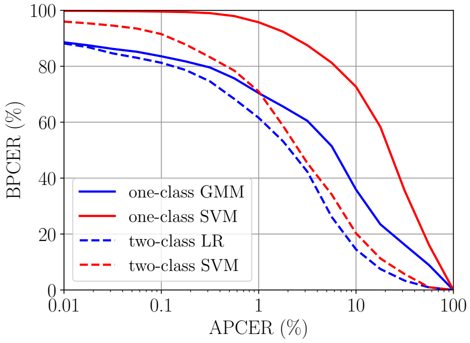

Figure 3. DET curves - *evaluation* set of the **grandtest** protocol of the Aggregated database. Four types of classifiers applied to IQM features of the facial region.

responds to generative methods [14], and one-class SVM with RBF kernel, which is a non-generative method [14]. For one-class GMM the output score is a weighted log-probability, for one-class SVM the output is a confidence score as returned by LIBSVM [5]. In the case of one-class GMM the number of clusters is set to $K = 50$, further augmentation of K gives no significant gain in performance. All one-class or anomaly detection systems are trained using *only bona-fide* samples of the training set. Among *two-class* classifiers a Logistic Regression (LR) and SVM with RBF kernel are evaluated. For both classifiers, the output score is a probability of a sample being a real class. The two-class systems are trained using real and attack samples of the training set.

First, all classifiers are evaluated using *"grandtest"* protocol of the Aggregated database, corresponding to the scenario when all types of attacks are present among all subset of the database. The EER, $HTER$ values are summarized in Table 3, with DET curves given in Figures 2 and 3.

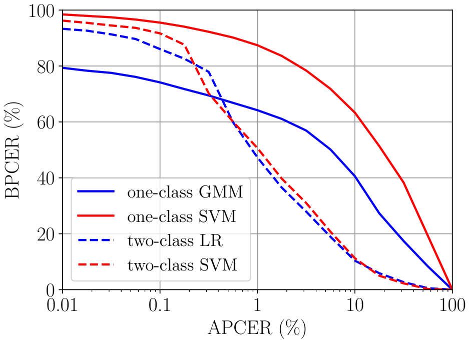

Figure 4. DET curves - *development* set of the **photo-photo-video** protocol of the Aggregated database. Four types of classifiers applied to IQM features of the facial region.

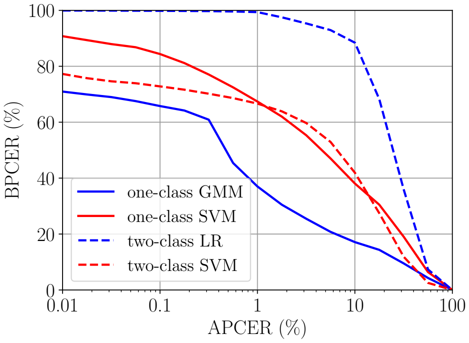

Figure 5. DET curves - *evaluation* set of the **photo-photo-video** protocol of the Aggregated database. Four types of classifiers applied to IQM features of the facial region.

As one can notice, two-class based systems in general perform better. Only for low $APCER$ values the performance of one-class GMM is on par with two-class systems, Figures 2, 3.

Next, the experiments using *"photo-photo-video"* and *"video-video-photo"* protocols are accumulated, addressing the case of unseen types of attacks in the evaluation set. For the *development* set, Figures 4 and 6, containing the same types of attacks as in the training step, the performance of two-class approaches is superior in comparison with anomaly detectors. However, in the *evaluation* set with unseen types of attacks, Figures 5 and 7, the opposite behavior can be observed - one-class GMM outperforms two-class methods in a wide range of $APCER$ values. Moreover, the relations of EER and $HTER$ values are more stable for one-class systems as opposed to two-class approaches, Table 3. For *"photo-photo-video"* protocol the $HTER$ values even go down compared to EER for both one-class SVM and GMM. This demonstrates a bet-

	Protocol	grandtest		photo-photo-video		video-video-photo	
	Subset	”dev”	”eval”	”dev”	”eval”	”dev”	”eval”
	Error, %	EER	HTER	EER	HTER	EER	HTER
Algorithm	one-class GMM	19.3	20.8	22.1	**14.5**	13.5	29.8
	one-class SVM	28.1	34.8	35.5	24.3	18.2	39.5
	two-class LR	10.3	**11.9**	10.2	30.1	1.5	30.3
	two-class SVM	12.7	15.3	10.5	21.9	1.4	**24.9**

Table 3. EER, computed on development - ”dev” set, and $HTER$, computed on evaluation ”eval” set, values for various protocols of the Aggregated database. Best $HTER$ per protocol is highlighted in bold.

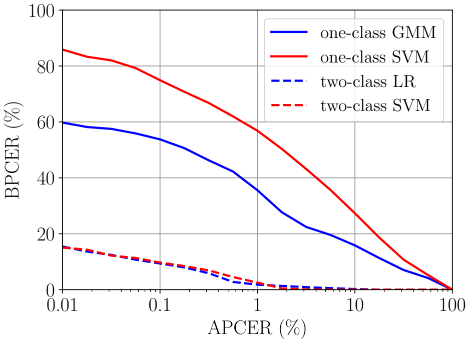

Figure 6. DET curves - *development* set of the **video-video-photo** protocol of the Aggregated database. Four types of classifiers applied to IQM features of the facial region.

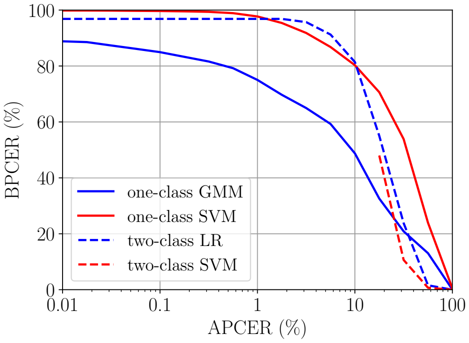

Figure 7. DET curves - *evaluation* set of the **video-video-photo** protocol of the Aggregated database. Four types of classifiers applied to IQM features of the facial region.

ter generalization properties of anomaly detectors against unseen types of attacks, which is a desirable feature in realistic scenarios of operation of facial PAD systems. Interestingly, that the best performing in the *”grandtest”* protocol, two-class LR classifier has the worst generalization properties against unseen types of attacks. This can be concluded from both high $HTER$ values, Table 3, and significant error growth in DET curves, Figures 5 and 7.

4. Conclusion

This paper addresses a problem of face anti-spoofing as an anomaly detection task. The work is motivated by the potentially wide diversity of spoofing attacks in realistic applications, and generalization issues specific to current nearly state-of-the-art PA detectors in the tests presenting unseen spoofs. Up to date, most of the research in the field was considering face PAD as a two-class classification problem: bona-fide samples versus attacks. To highlight the stated generalization issues we propose a new evaluation protocols designed to study the PAD algorithms under unseen attacks scenario. The protocols are made for introduced Aggregated database, which is composed of three publicly available sets: Replay-Attack, Replay-Mobile and MSU MFSD. Aggregated database unites the data collection efforts of various institutions making it a more challenging set, than any of its standalone components. It contains a wide variety of photo and replay attacks making a step towards more realistic evaluation scenario.

The lowest $HTER = 11.9\%$ for the *grandtest* protocol of the Aggregated database, among legacy systems based on two-class classifier, was obtained using IQM [12, 18] features and Logistic Regression. However, we then demonstrate poor generalization properties of this setup against unseen types of spoofs, giving 30.1% and 30.3% $HTER$ for the *photo-photo-video* and *video-video-photo* protocols of the Aggregated set. Moreover, the performance of the system drops significantly for lower APCER values, which is a desirable operation diapason in realistic applications.

In contrast, our proposed IQM and one-class GMM based PA detector generalizes better for *photo-photo-video* and *video-video-photo* protocols of the Aggregated database. The GMM anomaly detector is trained using only bona-fide samples present in the training set. The IQM-GMM system outperforms all other observed combinations in the wide range of APCER values. Only in the high APCER region, which is usually not of interest in realistic applications, the performance of two-class systems getting comparable with the proposed setup. Generally, though, it can be concluded that neither of the observed systems perform well enough in the proposed evaluation scenario, and there is a clear need to continue research in this direction.

Acknowledgments

This work was developed within the framework of the Biometric Authentication with a Timeless Learner (BATL) project, funded by US Intelligence Advanced Research Projects Activity (IARPA) agency, ODIN program.

References

[1] F. Alegre, A. Amehraye, and N. Evans. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, Sept 2013. 2

[2] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: A public database and a baseline. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–7, Oct 2011. 4

[3] S. R. Arashloo and J. Kittler. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *2017 International Joint Conference on Biometrics (IJCB)*, 2017. 2

[4] S. R. Arashloo, J. Kittler, and W. Christmas. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE Access*, 5:13868–13882, 2017. 2, 3

[5] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011. 5

[6] 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)*, pages 1–7, Sept 2012. 1, 2, 3, 4

[7] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel. The replay-mobile face presentation-attack database. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7, Sept 2016. 2, 3

[8] T. de Freitas Pereira, A. Anjos, J. M. D. Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *2013 International Conference on Biometrics (ICB)*, pages 1–8, June 2013. 2

[9] 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*, 39(1):1–38, 1977. 3

[10] Y. Ding and A. Ross. An ensemble of one-class svms for fingerprint spoof detection across different fabrication materials. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, Dec 2016. 2

[11] T. d. Freitas Pereira, J. Komulainen, A. Anjos, J. M. De Martino, A. Hadid, M. Pietikäinen, and S. Marcel. Face liveness detection using dynamic texture. *EURASIP Journal on Image and Video Processing*, 2014(1):2, Jan 2014. 1

[12] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Transactions on Image Processing*, 23(2):710–724, Feb 2014. 1, 2, 4, 6

[13] A. Hadid. Face biometrics under spoofing attacks: Vulnerabilities, countermeasures, open issues, and research directions. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 113–118, June 2014. 1

[14] J. Kittler, W. Christmas, T. de Campos, D. Windridge, F. Yan, J. Illingworth, and M. Osman. Domain anomaly detection in machine perception: A system architecture and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):845–859, May 2014. 3, 5

[15] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, pages 41.1 – 41.12. BMVA Press, 09 2015. 1

[16] U. Scherhag, R. Raghavendra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch. On the vulnerability of face recognition systems towards morphed face attacks. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, April 2017. 1

[17] Y. Sun, D. Liang, X. Wang, and X. Tang. DeepID3: Face Recognition with Very Deep Neural Networks. *arXiv preprint arXiv:1502.00873*, 2015. 1

[18] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, April 2015. 1, 2, 3, 4, 6

[19] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 26–31, March 2012. 2