Technical University of Denmark,
02238 Biometric Systems - Individual Term Paper

Evaluating Different Facial Regions for Presentation Attack Detection via Deep Learning

Nina Weng[1]

Abstract: With the comprehensive adoption of face recognition technique, Presentation Attack (PA) is considered the potential security threat. Presentation Attack Detection (PAD) is now widely researched to address this issue. However, most of the PAD approaches utilized the whole face, leading to degradation over detection performance. In this work, we analyze the performance of different facial regions in PAD. First, we introduce an algorithm for facial regions extraction, extracting 15 facial regions, including particular facial components, local facial regions, and global facial regions. Then a feature-level fusion CNN is implemented for PAD. We perform the task on two databases and then analyze the results based on Detection Error Tradeoff (DET) curve and Equal Error Rate (EER). The results show that the left middle face is the most efficient region for single region PAD and the combination of left middle face and mouth for region fusion PAD.

Keywords: Presentation attack detaction, facial regions, biometrics, facial components

1 Introduction

From unlocking the mobile phone to the access control for high-secret places, face recognition is now extensively applied in today's world. With the high accuracy and low intrusiveness, the usage for face recognition has been extended from the field of authentication to broader aspects, such as criminal tracking [El17] and mobile commerce payment [Du18].

The high level of development for face recognition in a broad range of security systems brings a new and urgent problem: how to discern the artefact from liveness. Among various kinds of attacks, Presentation Attack (PA) is one of the most common and easiest to implement attack means which executes on the sensor level. As shown in Figure 1, the typical PAs use the photo, the videos, or a 3D mask of the targeted person [MNL14].

Several face Presentation Attack Detection (PAD) methods have been proposed to address the potential security threats. From the hardware perspective, a camera contained the depth or thermal information, such as the light field camera [RRB15] or near-infrared (NIR) [BM17], directs a possible solution for the PAD problem. On the other hand, the software-based methods which are more flexible for facilitating utilize the hand-crafted features such as Local Binary Patterns (LBP) [XA18] [PQL20], Histogram of Oriented Gradients (HOG) [Ag17], and Local Phase Quantization (LPQ) [Ra18] with Support Vector Machine (SVM) to perform the binary classification. With the development of deep learning, approaches based on the neural network are proposed, for example, the fine-

[1] Technical University of Denmark, s202997@student.dtu.dk

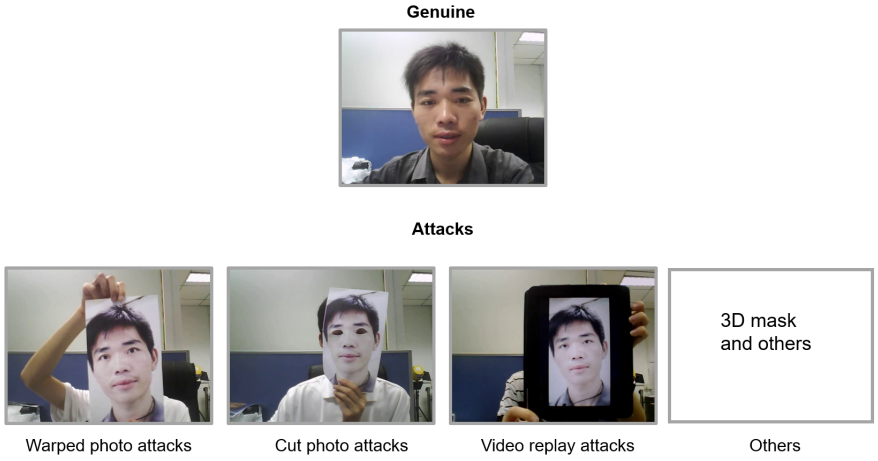


Fig. 1: The examples of Presentation Attack (PA), which are from CASIA Face Anti-Spoofing database [Zh12]. The upper row is the genuine target, while the bottom images are attacks. The attacks include warped photo attack, cut photo attack, and video replay attack

tuned ImageNet pre-trained CNN [YLL14] and two-stream CNN-based with patches and automatically extracted depth information [At17].

For most of the PAD methods, the whole face images are employed, leading to the potential degradation over detection performance. Besides, for some attack methods, the artefacts might replace the partial face to create the illusion of dynamic face, for instance, the cut photo attacks in CASIA Face Anti-Spoofing database [Zh12], which is shown in Figure 1. Furthermore, under non-laboratory scenes, some face components could be unclear due to the poor lighting or facial dressing (e.g., wired glasses, stickers on the side of the face), which results in the analysis restricted on parts of the face.

Based on the scenarios we mentioned above, analyzing the importance of different facial regions in presentation attack detection is thus meaningful. To our best knowledge, there is no previous research related to the facial regions to presentation attack detection yet, leaving this problem to be more challenging and demanding.

In this report, we explore the performance of different facial regions in the PAD task. Firstly, a robust algorithm for facial regions extraction is proposed. This algorithm is able to work under non-lab scenes, with the extraction of 15 regions from the particular regions (e.g., left eye) to the local regions (e.g., right middle face). After that, a Convolutional Neural Network (CNN) with feature-level fusion is built, allowing training and prediction of the single region and the combination of regions. Finally, the experiments are conducted on two databases, followed by the result analysis, including the effect of data representation for video-type data, the comparison of performance for different facial regions, and the facial regions fusion study.

The paper is organized as follows: in Section 2 the proposed method is well explained, including the algorithm for facial regions extraction and CNN-based feature-level fusion PAD method; the experimental setup is introduced in Section 3 followed by the result analysis and discussion in Section 4. Finally we conclude the report in Section 5.

2 Proposed Method

In this section, the method we use for this project is detailed introduced. Figure 2 presents the overall framework of the proposed method. For data representation, firstly, we extract frames from video-type data with a certain number of frames. The normalization process is applied based on the token face image geometric standard from ISO/IEC 19794.5:2011 [ISO11a]. With the normalized faces, the extraction of the facial region is performed as followed. We proposed a feature-level fusion CNN method for presentation attack detection, which could take one or multi regions as input.

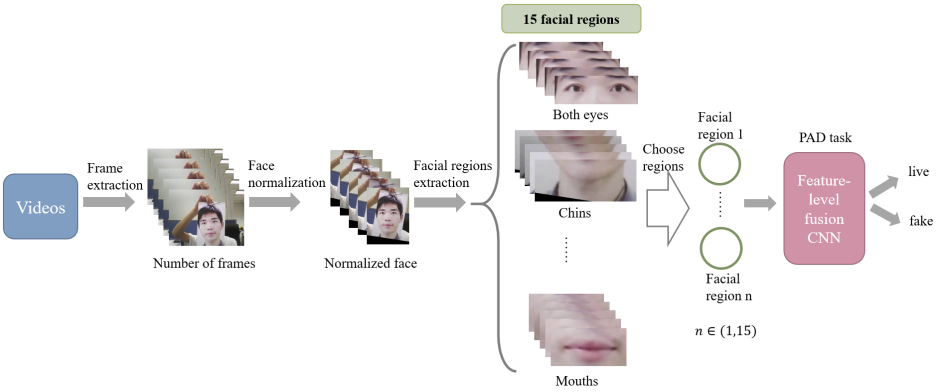


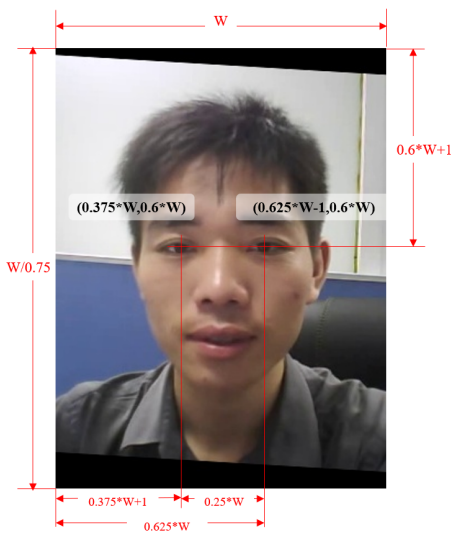
Fig. 2: The framework of proposed method in this work. After extracting frames from the video data, the face normalization is conducted according to the token face image standard from ISO/IEC 19794.5:2011 [ISO11a]. Then the extraction of 15 facial regions is performed. The PAD task is implemented by a feature-level fusion CNN which takes both single and multi regions as input.

2.1 Face extraction and normalization

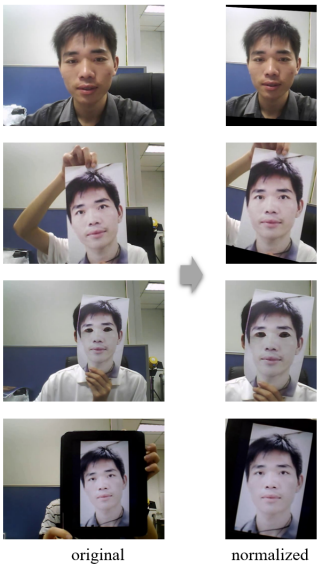
The very first step of this work is to represent the data from video to normalized face image. We are mainly focusing on the static analysis for PAD in this work; thus, the frames from the videos would be a good data representation. Considering only the slight changes occurs between different frames for PAD database, certain number of frames are used instead of all.

The face normalization is then conducted. The idea is that faces in different videos have different scales and locations, the normalized ones could be used as the uniform repre-

sentation regardless of their sources. The standard we use in this work is the token face image standard in ISO/IEC 19794.5:2011 [ISO11a]. To perform this, the face landmark detection [81] is applied in the first place; then, the affine operation is utilized according to inter-pupil distance and angle, finally, we cut the image with the pre-set size. After normalization, all face images are now with the width of 300 pixels, the height of 400 pixels, and the inter-pupil distance of 75 pixels. More details of the normalized face is shown in Figure 7a and with Figure 7b some examples are displayed for intuitive understanding.



(a) The token image standard from ISO/IEC 19794.5:2011 [ISO11a].



(b) Examples of normalized faces.

Fig. 3: Face normalization.

2.2 Facial regions extraction

The method of facial regions extraction used in this work is based on [To13]. Different from [To13], instead of using Luxand, we use a state-of-art facial landmark detector [81] which enables 81 landmark detection. This detection method is extended from dlib's 68 facial landmarks detection [Ki09]. Figure 4 shows the landmarks detection result in the left-top corner with the normalized face. Though 81 landmarks are detected, only 35 of them are used in our method. The determined landmarks, determined type, and the bounding box for each facial region are listed in Table 1. To extract a particular region, we average the location of determined landmark(s) and consider it as the position in determined type with the bounding box size. The results of facial regions extraction are shown in Figure 4.

The 15 facial regions contain: 1)precise facial components like left eye and mouth; 2) local regions in the face like both eyebrows, right middle face; 3) the global face like the

Tab. 1: Facial regions extraction parameters.

ID	Facial region	determined landmark(s)	determined type	bounding box
1	Chin	[8]	center	(75,181)
2	Left ear	[1,2]	center	(75,51)
3	Right ear	[14,15]	center	(75,51)
4	Left eyebrow	[17-21]	center	(51,75)
5	Right eyebrow	[22-26]	center	(51,75)
6	Both eyebrows	[21,22]	center	(51,151)
7	Left eye	[36-41]	center	(51,51)
8	Right eye	[42-47]	center	(51,51)
9	Both eyes	[39,42]	center	(51,151)
10	Full face (Face_ISOV)	[30]	center	(192,168)
11	Forehead	[21,22]	bottom center	(101,151)
12	Left middle face	[30]	right center	(173,106)
13	Right middle face	[30]	left center	(173,106)
14	Mouth	[61-63,65-67]	center	(51, 101)
15	Nose	[29]	center	(51,101)

full face. On the one hand, the more specific the region is, the fewer effects it receives from other parts of the face. On the other hand, the integration of face components could potentially reveal pieces of evidence for PAD.

2.3 CNN with feature-level fusion

Figure 5 shows the CNN structure we used in this work. For one region are chosen for PAD, only the resnet18 CNN [He16] pre-trained using ImageNet with the modification on the last FC layer (from 1000 output to 2 output) would be applied. For the fusion of multi regions PAD task, firstly, the pre-trained resnet18 CNN is implemented for each region, and the output (a vector with length 1000) could be considered a feature of that region. By concatenating all regions' features, we will get a vector with the length of $n * 1000$. After one more FC layer and the softmax layer, we will finally get the probabilities of being the bona fide or the attack, respectively.

The structure of resnet18 used in this work is presented in Table 2.

3 Experimental Setup

3.1 Dataset

In this work, two databases are used to evaluate regions or the combination of regions for PAD:

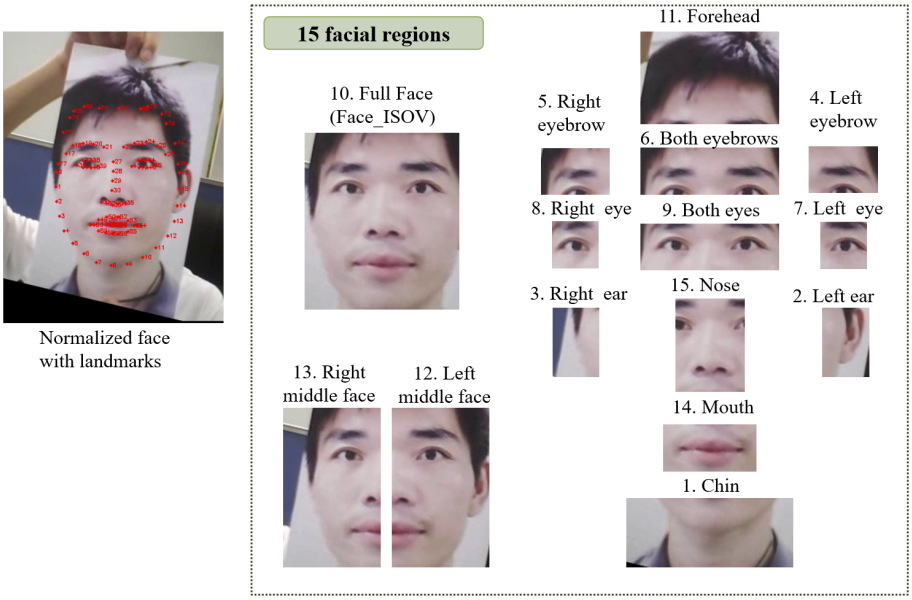


Fig. 4: The 15 facial regions extracted from the normalized face based on landmarks.

CASIA Face Anti-Spoofing database[Zh12]:There are 50 subjects in this database. For each subject, there are 12 videos with the combination of video quality(low, mid, high) and PAI species (warped photo attack, cut photo attack, and video replay attack). We extract 1 and 5 frames from each video as the representation of data.

REPLAY-MOBILE Subset[Co16]: The full REPLAY-MOBILE database 1190 video clips of printed attacks, photo replay attacks, and video replay attacks of 40 subjects under different lighting conditions. In this work, only a subset of the REPLAY-MOBILE database is used, containing 312 samples for training and 302 samples for testing. For bona fide samples, two kinds of capture devices and five types of lighting are included. For attacks, there are different combinations over display means and lighting conditions with the overall count of 16 types of attack for each subject.

3.2 Evaluation Metrics

The results of all experiments are analyzed under the standard from ISO/IEC 30107-3 [ISO11b] for biometric PAD:

- Attack Presentation Classification Error Rate (APCER), which is defined as the proportion of attack presentations wrongly classified as bona fide presentations.
- Bona Fide Presentation Classification Error Rate (BPCER), which is the proportion of bona fide presentations misclassified as attack presentations.

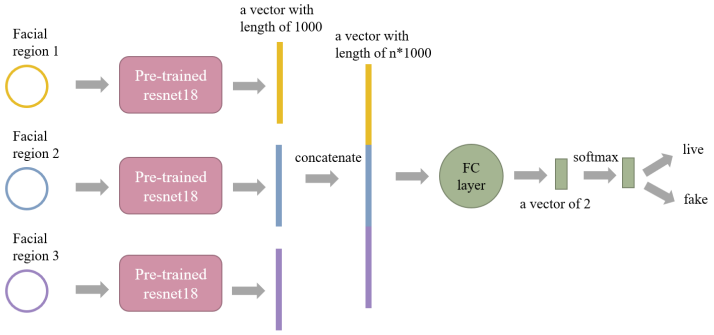


Fig. 5: The feature-level fusion CNN structure.

Tab. 2: Resnet18 structure used in this work

layer name	output size	Resnet-18
-(input)	256x256x3	-
conv1	128x128x64	7x7,64,stride2
conv2_x	64x64x64	3x3 max pool, stride 2 [3x3,64] [3x3,64] x2
conv3_x	32x32x128	[3x3,128] [3x3,128] x2
conv4_x	16x16x256	[3x3,256] [3x3,256] x2
conv5_x	8x8x512	[3x3,512] [3x3,512] x2
average pool	1x1x512	8x8 average pool
fully connected	1000 (for one region: 2)	512x1000 fully connection
softmax	-	-

Apart from these two, we also report the DET curves as well as Equal Error Rate (EER) in the experimental results.

3.3 Feature-level fusion CNN hyper-parameters

For the training process of feature-level fusion CNN, the batch size is set as 32; the number of epochs is set as 50 for single region PAD task and 100 for multi regions PAD task; learning rate is set as $1e-5$ for single region PAD and $1e-6$ for multi regions task.

4 Experimental Results

4.1 Effect of data representation for video clips

For video-like data, the data representation is of great importance and interest. While detailed data representation might improve the performance on the later task, it would also increase the burden on data pre-processing or the training process. The vital and sufficient information extraction is an art of trade-off.

In this work, we extract 1 and 5 frames per video for CASIA database respectively, and then conduct the single region PAD for both. The DET curves are showed in Figure 6 and the EER (%) for each regions are presented in Table 3.

It could be observed that with the increase of frames extracted from each video, the EERs decrease for all facial regions except the forehead. This finding also reveals from the DET curves, for data with 5 frames per video, the DET curves are more adjacent to the left-bottom corner.

4.2 PAD on different facial regions

As mentioned in Section 1, one of the key problems that we'd like to answer is that which facial region performs best in PAD task. Table 3 shows the Equal Error Rate for all face regions including the normalized face with different datasets. There are some key observations from this table: 1) For different dataset, the rank for *best facial region for PAD* might be different as well; 2) the local region seems to work better than the particular face components, not to mention the full face and the normalized face ; 3) among all particular facial areas, for CASIA dataset, *forehead* and *nose* seems to outperform the others remarkably; 4) for REPLAY-MOBILE dataset, the performance of the right side of face exceeds the left side of the face. What worth mentioning is that in the REPLAY-MOBILE dataset, one kind of lighting control, namely *'lateral'* provides that light from the left side, making the left side of the face is more bright, while the right side of the face is darker. This finding might be related to the shadows area in the face in lighting conditions, indicating that lighting for PAD might result in a better performance in the security system.

In Figure 6 the DET curves for both database with all face regions (except the ones with 0% EER) are displayed. Apart from the similar finding from Table 3, there are also some interesting stories behind those curves. According to the result from the CASIA database with 5 frames extracted per video, if the target system required low BPCER and has high tolerance with APCER (e.g. gym), then *forehead* would be a good choice compared with other regions; in contrary, if the target system required low APCER and has high tolerance with BPCER (e.g. top-secret lab), then *right middle face* would be the top choice.

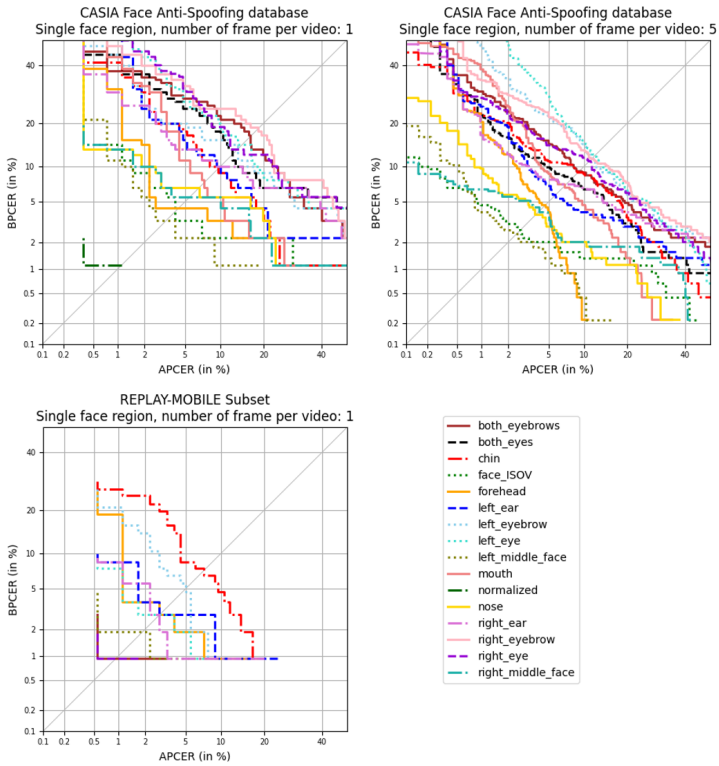


Fig. 6: DET curves over the single region PDA with different datasets and the different number of frames extracted pre video. Notice that some DET curves of facial regions or normalized faces are not shown in the plot due to 0.00% of EER.

4.3 Facial regions fusion study

The other key question that we would like to explore is that whether the combination of facial regions helps presentation attack detection compared with a single region. The idea comes from [To15], where the author discovers the combination of different regions of the human face in various forensic scenarios.

After choosing the best-performed region from the single region PAD, then we conduct experiments towards the combination of facial regions. Considering that for the REPLAY-MOBILE dataset, the lowest EER is zero, which has no more space for improvement, we choose the second-best region: *both eyebrows*. The DET curves are shown in Figure 7 and EERs are presented in Table 4. The results reveal that the combination of *mouth* and *left middle face* has the best performance among other combinations for CASIA database; for REPLAY-MOBILE, the combination with *forehead* and *mouth* wins others. Though some

Tab. 3: Benchmark with different face regions in terms of EER(%) for different datasets and the different number of frames extracted pre video. The top 3 regions with the smallest value are bolded in the table.

ID	Facial region	CASIA		REPLAY-MOBILE
		numf1	numf5	numf1
1	Chin	9.81	8.92	7.83
2	Left ear	10.00	5.97	2.80
3	Right ear	10.00	7.13	2.80
4	Left eyebrow	13.33	11.72	5.33
5	Right eyebrow	18.70	11.35	**0.75**
6	Both eyebrows	16.48	9.15	**0.75**
7	Left eye	14.44	12.02	2.80
8	Right eye	13.33	10.94	**0.75**
9	Both eyes	12.22	8.03	**0.75**
10	Full face (Face_ISOV)	**4.26**	**2.65**	**0.00**
11	Forehead	**4.44**	4.67	2.80
12	Left middle face	**3.33**	**2.43**	1.78
13	Right middle face	5.56	4.18	**0.00**
14	Mouth	7.59	7.05	1.78
15	Nose	6.67	**3.77**	1.50
-	Normalized face	1.11	0.00	0.00

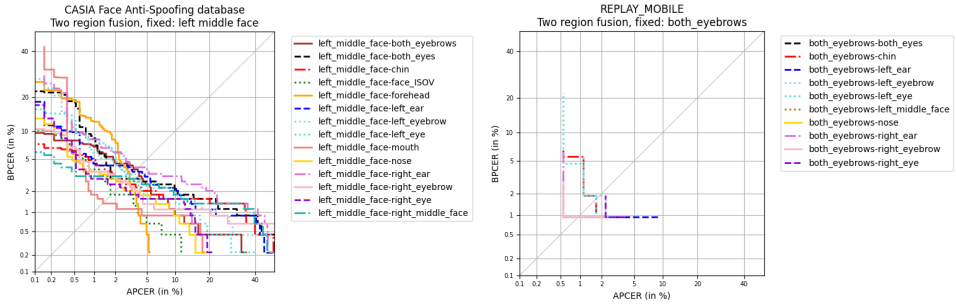
combinations do improve in EER, however, the results show no significant improvement for most of the combination.

4.4 Visualization on feature map

With deep learning methods, sometimes we are unsure what is the key factor inside the black box. Here we visualize the feature map to more intuitively understand what is learned by the CNN structure. Figure 8 shows the visualization of feature map for region *left middle face*. The four samples showed is the genuine, wrapped paper attack, cut paper attack, and video replay attack, respectively. It could be seen that for artefacts, the focus area is located around the upper left corner of the eye, whereas the genuine samples concentrate on the nose and left-side jaw instead. One possible reason for that is, for the wrapped paper attack, the upper left corner of the eye is very close to the place where it is used to hold the paper; and for the cut paper attack, the same satiation occurred again, furthermore the cut edge on paper is also around that location.

5 Conclusion

In this work, we propose a framework for extracting facial regions and evaluating them for presentation attack detection. A robust algorithm about facial regions extraction is first



(a) The 2 region fusion for CASIA database, fixed region: left middle face.

(b) The 2 region fusion for REPLAY-MOBILE database, fixed region: both$_e$*yebrows*.

Fig. 7: DET curves for facial regions fusion study

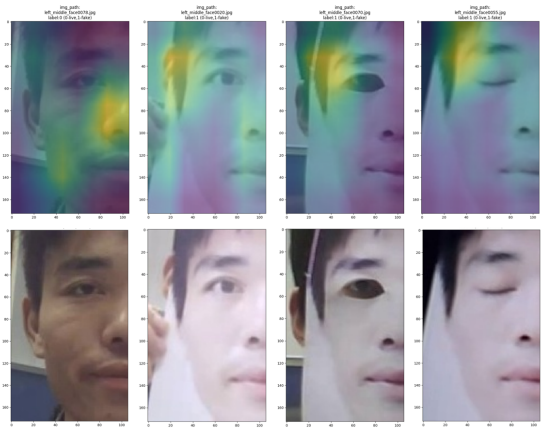


Fig. 8: The visualization of feature map for facial region *left middle face*. The 4 samples are genuine, wrapped paper attack, cut paper attack and video replay attack respectively.

introduced, which produces 15 facial regions, including particular facial components like eyes and mouth, local facial regions like half of the face, and the global facial regions as the full face. Then, a feature-level fusion CNN is implemented for PAD. We analyze the DET curves and compute the EER for each single region and the combination of regions on two databases. The result shows that the local regions, such as the left middle face, work much better than the particular facial component (e.g., left eye). Among the particular facial components, the nose and forehead are the most efficient regions to distinguish artefacts from liveness. We also exam the combination of facial regions, which tells us the combination of mouth and left middle face works the best with the EER of 1.46%.

Tab. 4: Facial regions fusion study result with EER (%). The values are bolded when it is lower than EER from fixed single region & lower than itself EER with single region.

ID	Facial region	CASIA fixed region: left middle face	REPLAY-MOBILE fixed region: both eyebrows
0	*only fixed*	2.43	0.75
1	Chin	2.65	1.78
2	Left ear	3.55	0.75
3	Right ear	3.99	0.75
4	Left eyebrow	3.32	1.78
5	Right eyebrow	2.65	0.75
6	Both eyebrows	3.77	-
7	Left eye	3.25	1.78
8	Right eye	2.39	1.78
9	Both eyes	2.87	0.75
10	Full face (Face_ISOV)	**1.75**	0.00
11	Forehead	3.10	**0.00**
12	Left middle face	-	0.75
13	Right middle face	2.87	00.00
14	Mouth	**1.46**	**0.00**
15	Nose	2.43	0.75

References

[81] 81 Facial Landmarks Shape Predictor. https://github.com/codeniko/shape$_predictor_81_face_landmarks.author$: *codeniko*.

[Ag17] Agarwal, Akshay; Yadav, Daksha; Kohli, Naman; Singh, Richa; Vatsa, Mayank; Noore, Afzel: Face presentation attack with latex masks in multispectral videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 81–89, 2017.

[At17] Atoum, Yousef; Liu, Yaojie; Jourabloo, Amin; Liu, Xiaoming: Face anti-spoofing using patch and depth-based CNNs. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, pp. 319–328, 2017.

[BM17] Bhattacharjee, Sushil; Marcel, Sébastien: What you can't see can help you-extended-range imaging for 3d-mask presentation attack detection. In: 2017 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, pp. 1–7, 2017.

[Co16] Costa-Pazo, Artur; Bhattacharjee, Sushil; Vazquez-Fernandez, Esteban; Marcel, Sebastien: The replay-mobile face presentation-attack database. In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, pp. 1–7, 2016.

[Du18] Du, Meiyan: Mobile payment recognition technology based on face detection algorithm. Concurrency and Computation: Practice and Experience, 30(22):e4655, 2018.

[El17] Elrefaei, Lamiaa A.; Alharthi, Alaa; Alamoudi, Huda; Almutairi, Shatha; Al-rammah, Fatima: Real-time face detection and tracking on mobile phones for criminal detection. In: 2017 2nd International Conference on Anti-Cyber Crimes (ICACC). pp. 75–80, 2017.

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778, 2016.

[ISO11a] : Information Technology - Biometric Data Interchange Formats - Part 5: Face Image Data. Standard, International Organization for Standardization, Geneva, CH, March 2011.

[ISO11b] : Information technology - biometric presentation attack detection – Part 3: testing and reporting (2017). Standard, International Organization for Standardization, Geneva, CH, March 2011.

[Ki09] King, Davis E: Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, 10:1755–1758, 2009.

[MNL14] Marcel, Sébastien; Nixon, Mark S; Li, Stan Z: Handbook of biometric anti-spoofing, volume 1. Springer, 2014.

[PQL20] Peng, Fei; Qin, Le; Long, Min: Face presentation attack detection based on chromatic co-occurrence of local binary pattern and ensemble learning. Journal of Visual Communication and Image Representation, 66:102746, 2020.

[Ra18] Raghavendra, Ramachandra; Venkatesh, Sushma; Raja, Kiran B; Wasnik, Pankaj; Stokkenes, Martin; Busch, Christoph: Fusion of multi-scale local phase quantization features for face presentation attack detection. In: 2018 21st International Conference on Information Fusion (FUSION). IEEE, pp. 2107–2112, 2018.

[RRB15] Raghavendra, Ramachandra; Raja, Kiran B; Busch, Christoph: Presentation attack detection for face recognition using light field camera. IEEE Transactions on Image Processing, 24(3):1060–1075, 2015.

[To13] Tomé, Pedro; Blázquez, Luis; Vera-Rodriguez, Rubén; Fierrez, Julián; Ortega-García, Javier; Expósito, Nicomedes; Lestón, Patricio: Understanding the discrimination power of facial regions in forensic casework. In: 2013 International Workshop on Biometrics and Forensics (IWBF). IEEE, pp. 1–4, 2013.

[To15] Tome, Pedro; Fierrez, Julian; Vera-Rodriguez, Ruben; Ortega-Garcia, Javier: Combination of face regions in forensic scenarios. Journal of forensic sciences, 60(4):1046–1051, 2015.

[XA18] Xiong, Fei; AbdAlmageed, Wael: Unknown presentation attack detection with face rgb images. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–9, 2018.

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

[Zh12] Zhang, Zhiwei; Yan, Junjie; Liu, Sifei; Lei, Zhen; Yi, Dong; Li, Stan Z: A face antispoofing database with diverse attacks. In: 2012 5th IAPR international conference on Biometrics (ICB). IEEE, pp. 26–31, 2012.