FACE SPOOFING DETECTION BY FUSING BINOCULAR DEPTH AND SPATIAL PYRAMID CODING MICRO-TEXTURE FEATURES

Xiao Song, Xu Zhao*, Tianwei Lin

Key Laboratory of System Control and Information Processing MOE
Department of Automation, Shanghai Jiao Tong University

ABSTRACT

Robust features are of vital importance to face spoofing detection, because various situations make feature space extremely complicated to partition. Thus in this paper, two novel and robust features for anti-spoofing are proposed. The first one is a binocular camera based depth feature called Template Face Matched Binocular Depth (TFBD) feature. The second one is a high-level micro-texture based feature called Spatial Pyramid Coding Micro-Texture (SPMT) feature. Novel template face registration algorithm and spatial pyramid coding algorithm are also introduced along with the two novel features. Multi-modal face spoofing detection is implemented based on these two robust features. Experiments are conducted on a widely used dataset and a comprehensive dataset constructed by ourselves. The results reveal that face spoofing detection with the fusion of our proposed features is of strong robustness and time efficiency, meanwhile outperforming other state-of-the-art traditional methods.

Index Terms— Face spoofing detection, binocular depth, template face registration, spatial pyramid coding, micro-texture feature

1. INTRODUCTION

Recently, face spoofing detection develops into a deeply concerned research topic in computer vision field. Most of anti-spoofing methods existed can be categorized into three chief varieties: physiological sign based approaches, texture based approaches and illumination peculiarity based approaches. A good survey of research against spoofing attacks can be found in [1, 2]. Physiological sign based methods aim at capturing biometric motions such as eye blinking [3, 4], mouth movements [5] and the holistic facial motions [6, 7], etc. However, most of the methods mentioned above are performed in face image sequences only, and may fail when attacked by movable 3D models or videos of live people.

Another generally used facial cues for spoofing detection is illumination peculiarity [8, 9]. For instance, Zhang *et al*

[10] analyze the different multi-spectral reflectance distributions between genuine and fake faces using Lambertian model. Some researchers merely utilize the illumination cue from a single image [11, 12]. For instance, Tan *et al* [11] employ Retinex-based method to extract illumination reflectance feature for classification. However, these approaches always need extra devices and may fail when attacking photos or prints are of high qualities.

In [13], it is demonstrated that local micro-texture is a useful cue when attacked by recaptured images. Maatta *et al* [14] analyze micro-texture features with diverse operators. In [15], the same authors present a novel micro-texture descriptor called Multi-Scale Local Binary Patterns (MSLBP). Freitas *et al* [16] propose LBPTOP operator by fusing space information with time information. However, the micro-texture feature is low-level thus they are sensitive to intense illumination change and prints or photos of high qualities.

However, most above-mentioned methods are based on low-level descriptors and barely exploit the information derived from rigid structure of live face. The main contribution of our work is that two novel features are proposed for multi-modal classification. Binocular camera system is adopted to calculate original depth value. Every detected facial key point is augmented with the third dimension of depth, and then transformed by our proposed registration transformation to match corresponding key point in template face for several iterations. Afterwards Template Face Matched Binocular Depth feature vector is constructed. MSLBP [15] descriptor is applied to each pixel in facial region and Bag of Visual Word (BOVW) code [17] per pixel is obtained using pretrained codebooks. Then our spatial pyramid coding method is implemented, where weighted and normalized BOVW code histograms, as well as matching-degree vectors matching with two average intra-class faces from corresponding sub-regions, are concatenated as Spatial Pyramid Coding Micro-Texture feature vector. The system architecture is shown in Fig. 1.

We set up nonparallel dual cameras and conduct stereo calibration, obtaining two images at the same time, from which TFBD feature and SPMT feature are extracted. Each feature vector is individually fed to a nonlinear SVM classifier and score-level fusion of two individual SVM outputs determines classification result.
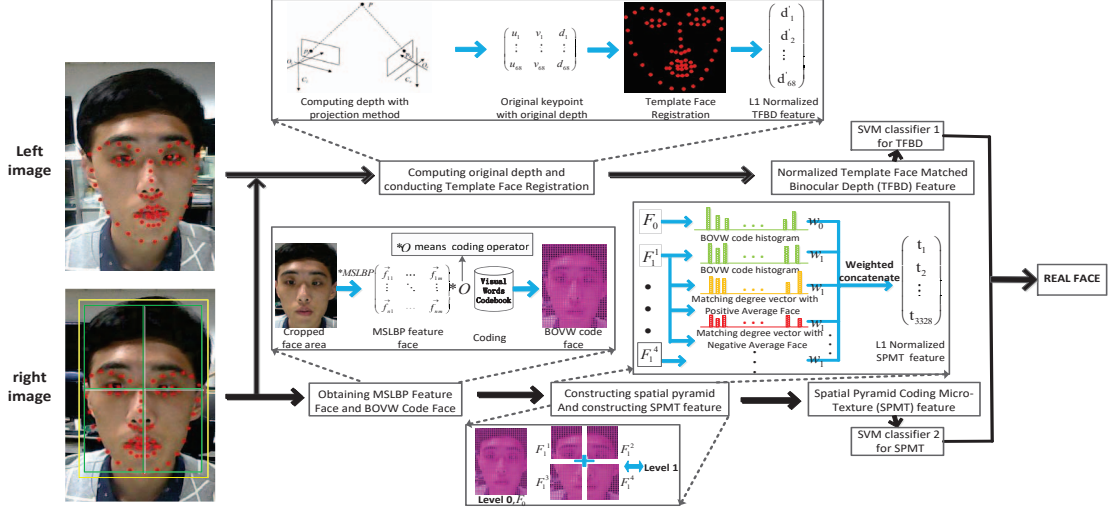
Fig. 1. The architecture of face spoofing detection with our proposed features

2. SPOOFING DETECTION WITH PROPOSED FEATURES

2.1. Template Face Matched Binocular Depth Feature

The Constrained Local Models (CLM) [18] is adopted to locate 68 facial key points and obtain their pixel coordinates in both left and right images.

E is defined as rotation matrix and V represents translation vector. $M_r = \{f_{xr}, c_{xr}, f_{yr}, c_{yr}\}$, M_l are defined as intrinsic matrices of right and left cameras respectively. $p_l = [u_l, v_l, 1]^T$, $p_r = [u_r, v_r, 1]^T$ are homogeneous pixel coordinates of certain keypoint in left image and right image. $m = M_l \begin{bmatrix} E & V \\ 0^T & 1 \end{bmatrix}$. According to pinhole camera model, original depth value of certain facial keypoint d is obtained:

$$d = \frac{B_{12}b_2 - B_{22}b_1}{\frac{u_r - c_{xr}}{f_{xr}}(B_{12}B_{21} - B_{11}B_{22}) + (B_{12}B_{23} - B_{22}B_{13})} \quad (1)$$

in which $B_{1j} = m_{1j} - m_{3j}u_l, B_{2j} = m_{2j} - m_{3j}v_l, b_1 = m_{34}u_l - m_{14}, b_2 = m_{34}v_l - m_{24}$ are intermediate variables.

Three dimensional abstract keypoint is defined for registration operation where the first and second dimensions are pixel coordinate in the right image and the third dimension is normalized depth value which powerfully reflects stereo structure of face. Each face to be detected can be denoted by a set with 68 abstract keypoints: $\{p_j | p_j = [x_j, y_j, d'_j]^T, 1 \leq j \leq 68\}$. A set of standard abstract keypoints called Template Face T is needed: $T = \{T_j | T_j = [T_x^j, T_y^j, T_d^j]^T, 1 \leq j \leq 68\}$. Template face is obtained before training. We select 20 image pairs collected from 5 different subjects with moderate distance away from cameras. Cameras are usually placed directly in front of their faces while pictures are shot.

x_i^j, y_i^j are defined as x and y values of the jth abstract keypoint in the ith right picture among 20 selected pairs. d_i^j and d'^j_i represent original depth and normalized depth: $d'^j_i = d_i^j - (\sum_j d_i^j)/68$. Template Face is obtained as follows:

$$T_x^j = \sum_i x_i^j/20, \ T_y^j = \sum_i y_i^j/20, \ T_d^j = \sum_i d'^j_i/20 \quad (2)$$

Single-round registration paradigm is defined as $\hat{p}_j = s \cdot R \times p_j + t$, where s represents scaling factor, R, t indicate rotation matrix and translation vector for abstract keypoint. Every round of registration transformation seeks for optimal parameters with minimum registration error:

$$(s^*, R^*, t^*) = argmin \sum_{j=1}^{68} \| T^j - s \cdot R \times p_j - t \|^2 \quad (3)$$

Absolute orientation using unit quaternions algorithm [19] is adopted to solve Eq.(3). Every abstract key point $\hat{p}_j$ can be solved with (s^*, R^*, t^*) after single round of registration.

An optimized iteration method is proposed based on Iterative Closest Point algorithm. We collect 20 abstract keypoints with minimum errors from current round of registration as parameters, to search optimal parameters (s^*, R^*, t^*) in the next round. After 20 rounds of registraion, 68 depth values extracted from 68 abstract keypoints is concatenated as 68 dimensional TFBD Feature vector.

2.2. Spatial Pyramid Coding Micro-Texture Feature

Anti-spoofing method based on only TFBD feature may fail when three dimensional structures of fake faces are highly similar to real faces. Hence, high-level micro-texture feature is needed to fuse with binocular depth feature.

A cascade detect model [20] is implemented to detect the

face region in right image. Before cropped, the detected face area should be expanded because it is demonstrated in [13] that most discriminative areas locate in marginal areas of face. Afterwards, the cropped face area is normalized into a 64×72 pixel gray-scale image.

Basic MSLBP operator using circular neighborhood introduced in [15] is adopted. The notation $LBP_{P,R}^u$ indicates that P sampled pixels on circle with a radius of R are compared with central pixel of neighborhood and uniform pattern [15] is adopted to transform original LBP value, reducing the amount of labels for $LBP_{P,R}$.

A MSLBP operator with a capacity of 3 illustrated as $\{LBP_{8,1}^u, LBP_{8,2}^u, LBP_{16,2}^u\}$ is applied per pixel in normalized face image. MSLBP Feature Face F_{mp}^u is defined as low-level texture descriptor. Each "pixel" in F_{mp}^u has three channels: $F_{mp}^u(x,y) = [LBP_{8,1}^u(x,y), LBP_{8,2}^u(x,y), LBP_{16,2}^u(x,y)]$. Values of three LBP used are 8-bit long after transformed with uniform pattern.

A medial-level texture descriptor called BOVW code face F_{bw} is introduced. A MSLBP Feature codebook with a capacity of 256 is obtained by K-means clustering algorithm before training and testing, where we select 3000 representative MSLBP feature faces from training set for clustering. Codebook can be notated as: $\{code_i | code_i = [code_i^1, code_i^2, code_i^3]\}$ for $1 \le i \le 256$. Afterwards BOVW coding algorithm in Eq. (4) is applied per "pixel" in MSLBP feature Face, obtaining BOVW code face with the size of 64×72:

$$F_{bw}(x,y) = argmin_k \sum_{n=1}^{3} \parallel code_k^n - F_{mp}^u(x,y)[n] \parallel^2 \quad (4)$$

Inspired by Spatial Pyramid Matching [21], a novel spatial pyramid coding algorithm is proposed. BOVW Code Face preserves original spatial layout thus it can be partitioned. In the first step, two-level pyramid is constructed on F_{bw} where level l means each spatial dimension is subdivided by factor 2^l for $l = 0, 1$. Let F_0^1 notate whole BOVW code face, F_1^i for $i = 1, 2, 3, 4$ notate the ith sub-region under level 1. Thus 5 sub-regions are obtained. The higher level of resolution is, the greater weight sub-region owns. Weight pyramid is adopted with $w_l = \frac{1}{2^{L-l}}$, where w_l notates partitioned sub-region's weight under level l and L represents maximal level of resolution. Each normalized BOVW code histograms constructed from corresponding sub-region is weighted by w_l.

In the second step, two datasets independent from training and test set are collected and named as positive and negative average sets, in which 2000 positive samples and 2000 negative samples are included respectively. Then average intra-class face is introduced:

$$A_{NorP} = \frac{1}{|\Omega_{NorP}|} \sum_{i \in \Omega_{NorP}} F_{bw}^i \quad (5)$$

where Ω_{NorP} represents positive or negative average set. Thus Positive Average Intra-class Face A_P and Negative Average Intra-class Face A_N are obtained.

Matching-degree vector is introduced in Eq.(6), where

$1(\cdot)$ is indicator function, $\gamma = N, P$ indicates matching with positive or negative average face, l is resolution level, $c \in [0, 255]$ represents BOVW code value, $f_l^i(c) = \sum_{(x,y)} 1(F_l^i(x,y) = c)$, $a_{\gamma,l}^i(c) = \sum_{(x,y)} 1(A_{\gamma,l}^i(x,y) = c)$, i means ith sub-region, so each M is 256 dimensional.

$$M_{\gamma,l}^i(c) = \begin{cases} w_l \times 1 & f_l^i(c) = 0 \& \& a_{\gamma,l}^i(c) = 0 \\ w_l \times min(\frac{f_l^i(c)}{a_{\gamma,l(c)}^i}, \frac{a_{\gamma,l}^i(c)}{f_l^i(c)}) & others \end{cases} \quad (6)$$

Two-level pyramid is constructed on A_P and A_N, obtaining $A_{P,1}^i$ and $A_{N,1}^i$ under level 1 for $i = 1, 2, 3, 4$. For each sub-region F_1^i, two matching-degree vectors $M_{P,1}^i$ and $M_{N,1}^i$ matching with $A_{P,1}^i$ and $A_{N,1}^i$ respectively are computed and L-1 normalized.

Finally, 5 normalized and weighted BOVW code histograms as well as 8 normalized and weighted matching-degree vectors are concatenated as 3328 dimensional Spatial Pyramid Coding Micro-Texture feature vector.

2.3. Classification

Once 68 dimensional TFBD feature and 3328 dimensional SPMT feature are obtained, each feature vector is individually fed to a well trained nonlinear SVM classifier. Score-level fusion of two individual SVM outputs is adopted because effects of TFBD and SPMT features on spoofing detection are independent and of the same order.

3. EXPERIMENTS

3.1. Datasets and Setup

Publicly available NUAA Photograph Imposter Database [12] and CASIA-FASD Database [22] are employed. Considering no available public binocular camera based dataset for face spoofing detection, we construct our own dataset which is composed of 6000 image pairs captured with two web cameras. Each pair includes a picture taken by left camera and a picture taken by right camera simultaneously. Our machine is a modern PC with 8GB RAM and GTX960 graphics card.

Dual cameras need to be calibrated before computing depth value thus our dataset is sampled with two fixed and calibrated cameras with resolution 640×480, meanwhile our proposed algorithm is resolution tolerant because of its keypoint and ROI based property. 15 different subjects are involved and 200 image pairs are sampled from each person under different illumination condition. People are also required to raise head, lower head, spin face, sit with different position and varied distance away from cameras. Meanwhile 20 images of different faces with high definition are printed on A4 papers, other 5 faces are printed on photographic papers and another 5 faces are displayed on an ipad screen. 30 fake faces are obtained in total and 100 picture pairs are sampled from each fake face under different illumination

condition. For every fake face, we move it horizontally, vertically, back and front and rotate it in depth. Especially for those printed on papers or photos, we also bend them inward and outward.

Two individual SVM classifiers for TFBD feature and SPMT feature need to be trained respectively. 3200 image pairs (1600 positive pairs, 1600 negative pairs) are used for training of the TFBD SVM classifier, meanwhile other 2800 image pairs in our dataset form the test set for evaluating our two proposed features. Multiple images from three datasets mentioned above are used to train SPMT SVM classifier. Test set in NUAA database is used to solely test our proposed SPMT feature and make comparisons with other descriptors.

Table 1. Comparison of texture descriptors on NUAA dataset

Operator	LPQ [23]	Tan's [12]	Mslbp [15]	Yang's [13]	**SPMT (ours)**
Accuracy	0.870	0.881	0.928	0.975	**0.979**
AUC	0.931	0.941	0.977	0.992	**0.995**
EER	14.8%	13.9%	8.0%	2.2%	**2.0%**

Table 2. Comparison of texture descriptors on our dataset

Operator	LPQ [23]	Tan's [12]	Mslbp [15]	Yang's [13]	**SPMT (ours)**
Accuracy	0.851	0.860	0.916	0.937	**0.945**
AUC	0.914	0.922	0.969	0.982	**0.986**
EER	17.0%	15.9%	10.1%	6.5%	**5.0%**

Table 3. Evaluation of our proposed features on our dataset

Operator	Original depth	TFBD feature	SPMT feature	**SPMT+ TFBD**
Accuracy	0.871	0.920	0.945	**0.992**
AUC	0.933	0.972	0.986	**0.998**
EER	14.7%	8.2%	5.0%	**0.3%**

3.2. Experimental Results

The first experiment solely evaluates our proposed SPMT feature. Area Under Curve (AUC), Equal Error Rate (EER) and accuracy are adopted as assessment criteria. Comparisons are made with four powerful texture descriptors. For instance, MSLBP [15] is the state-of-the-art low level descriptor and Yang's component dependent descriptor [13] is a good medial-level descriptor. Optimal SVM parameters are used for each descriptor. Experiment is conducted on NUAA database and results are shown in Table 1. As can be seen, our SPMT descriptor outperforms slightly than Yang's, but not obvious. But in terms of time efficiency, component dependent descriptor runs at 4 fps while our SPMT descriptor runs at 10 fps which is 2.5 times faster.

In NUAA database, cameras always be in front of faces and the distance of the face away from cameras is always moderate. The second experiment is conducted on our own challenging dataset. The results are shown in Table 2. We can see that all descriptors perform worse but SPMT feature still
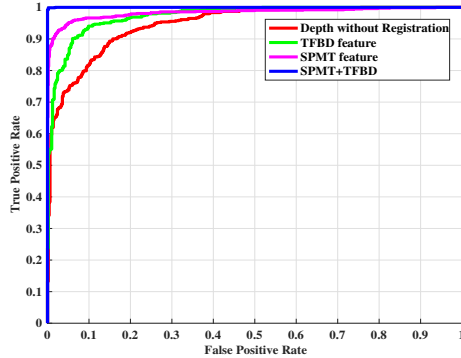

Fig. 2. ROC curves of our proposed feature descriptors evaluated on our dataset

performs best. For instance, component dependent descriptor needs to extract feature histograms from 4 components such as left eye, but when face is spun or far away from camera it is hard to locate components precisely. SPMT descriptor aims at excavating global intra-class similarity and spatial distribution of texture, thus sensitivity of feature is reduced.

The third experiment is also conducted on our own dataset. Our TFBD descriptor runs at 50 fps hence decision level fusion of two features are computational efficient. The results shown in Table 3 reveal that original depth feature is somewhat discriminative, however when face is increasingly far from cameras, relative differences of depth between different key points are shrinking. Performance also worsens when face in front of camera is spun because of the error caused by imprecise location of key points. After matched with template face, normalized binocular depth feature fully reflects stereo structure of face. 5% improvement in accuracy proves the effectiveness of TFBD feature. But TFBD feature has limitation in its sensitivity to stereo structure which is highly similar to real face. Hence SPMT feature is introduced and its effectiveness is proved by accuracy nearly 95%. However, limitation also exists in its sensitivity to intense illumination change and high definition print. Hence, the method of fusing TFBD feature with SPMT feature is adopted eventually. As can be seen, the accuracy is finally improved to 99.2% and AUC value is improved to 0.998. ROC curves of our proposed features evaluated on our own dataset are shown in Fig. 2.

4. CONCLUSION

Proposed TFBD feature highlights differences in three dimensional structures between genuine and fake faces, while the proposed SPMT feature highlights global intra-class similarities and spatial distribution of micro-texture. The multimodal detection based on these robust features shows comparative advantages on identifying fake faces displayed on photos, prints or videos with high definition. We believe our features can also be applied to other face recognition tasks.

5. REFERENCES

[1] Kristin Adair Nixon, Valerio Aimale, and Robert K Rowe, "Spoof detection schemes," in *Handbook of biometrics*, pp. 403–423. Springer, 2008.

[2] Gang Pan, Lin Sun, and Zhaohui Wu, *Liveness detection for face recognition*, INTECH Open Access Publisher, 2008.

[3] Lin Sun, Gang Pan, Zhaohui Wu, and Shihong Lao, "Blinking-based live face detection using conditional random fields," in *International Conference on Biometrics*. Springer, 2007, pp. 252–260.

[4] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[5] Klaus Kollreider, Hartwig Fronthaler, Maycel Isaac Faraj, and Josef Bigun, "Real-time face detection and motion analysis with application in liveness assessment," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 548–558, 2007.

[6] Klaus Kollreider, Hartwig Fronthaler, and Josef Bigun, "Non-intrusive liveness detection by face images," *Image and Vision Computing*, vol. 27, no. 3, pp. 233–244, 2009.

[7] Wei Bao, Hong Li, Nan Li, and Wei Jiang, "A liveness detection method for face recognition based on optical flow field," in *2009 International Conference on Image Analysis and Signal Processing*. IEEE, 2009, pp. 233–236.

[8] Youngshin Kim, Jaekeun Na, Seongbeak Yoon, and Juneho Yi, "Masked fake face detection using radiance measurements," *JOSA A*, vol. 26, no. 4, pp. 760–766, 2009.

[9] Allan Pinto, Helio Pedrini, William Robson Schwartz, and Anderson Rocha, "Face spoofing detection through visual codebooks of spectral temporal cubes," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4726–4740, 2015.

[10] Zhiwei Zhang, Dong Yi, Zhen Lei, and Stan Z Li, "Face liveness detection by learning multispectral reflectance distributions," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 436–441.

[11] Jiangwei Li, Yunhong Wang, Tieniu Tan, and Anil K Jain, "Live face detection based on the analysis of fourier spectra," in *Defense and Security*. International Society for Optics and Photonics, 2004, pp. 296–303.

[12] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *European Conference on Computer Vision*. Springer, 2010, pp. 504–517.

[13] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li, "Face liveness detection with component dependent descriptor," in *2013 International Conference on Biometrics (ICB)*. IEEE, 2013, pp. 1–6.

[14] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen, "Face spoofing detection from single images using texture and local shape analysis," *IET biometrics*, vol. 1, no. 1, pp. 3–10, 2012.

[15] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *Biometrics (IJCB), 2011 international joint conference on*. IEEE, 2011, pp. 1–7.

[16] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel, "Lbp-top based countermeasure against face spoofing attacks," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 121–132.

[17] Gabriella Csurka, Christopher Dance, and Lixin Fan, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*. Prague, 2004, vol. 1, pp. 1–2.

[18] David Cristinacce and Tim Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.

[19] Berthold KP Horn, "Closed-form solution of absolute orientation using unit quaternions," *JOSA A*, vol. 4, no. 4, pp. 629–642, 1987.

[20] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001*. IEEE, 2001, vol. 1, pp. I–511.

[21] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006*. IEEE, 2006, vol. 2, pp. 2169–2178.

[22] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li, "A face antispoofing database with diverse attacks," in *2012 5th IAPR International Conference on Biometrics (ICB)*. IEEE, 2012, pp. 26–31.

[23] Ville Ojansivu and Janne Heikkilä, "Blur insensitive texture classification using local phase quantization," in *International conference on image and signal processing*. Springer, 2008, pp. 236–243.