J. Vis. Commun. Image R. 38 (2016) 451–460

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Integration of image quality and motion cues for face anti-spoofing: A neural network approach ☆

Litong Feng [a,*], Lai-Man Po [a], Yuming Li [a], Xuyuan Xu [a], Fang Yuan [a], Terence Chun-Ho Cheung [b], Kwok-Wai Cheung [c]

[a] Department of Electronic Engineering, City University of Hong Kong, Hong Kong Special Administrative Region
[b] Department of Information Systems, City University of Hong Kong, Hong Kong Special Administrative Region
[c] Department of Computer Science, Chu Hai College of Higher Education, Hong Kong Special Administrative Region

ARTICLE INFO

Article history:
Received 29 June 2015
Revised 5 January 2016
Accepted 20 March 2016
Available online 1 April 2016

Keywords:
Face anti-spoofing
Neural network
Feature fusion
Shearlet
Dense optical flow

ABSTRACT

Many trait-specific countermeasures to face spoofing attacks have been developed for security of face authentication. However, there is no superior face anti-spoofing technique to deal with every kind of spoofing attack in varying scenarios. In order to improve the generalization ability of face anti-spoofing approaches, an extendable multi-cues integration framework for face anti-spoofing using a hierarchical neural network is proposed, which can fuse image quality cues and motion cues for liveness detection. Shearlet is utilized to develop an image quality-based liveness feature. Dense optical flow is utilized to extract motion-based liveness features. A bottleneck feature fusion strategy can integrate different liveness features effectively. The proposed approach was evaluated on three public face anti-spoofing databases. A half total error rate (HTER) of 0% and an equal error rate (EER) of 0% were achieved on both REPLAY-ATTACK database and 3D-MAD database. An EER of 5.83% was achieved on CASIA-FASD database.

1. Introduction

Face recognition has been wildly applied in user authentication systems due to the non-intrusive and natural interaction of face biometrics [1–3]. A high security requirement for face authentication is urgent, because only a photography, video replay, or 3D-mask can easily spoof a face recognition system to access secure information illegally [4]. Photos and videos of a valid user can be easily obtained, especially through social network. Hence the face anti-spoofing module is indispensable to a face authentication system besides the face recognition module. Recently, many attentions have been paid to countermeasures to facial spoofing attacks. A variety of face anti-spoofing algorithms were proposed based on different approaches [5]. Several public face anti-spoofing databases were established [6–8]. And competitions on countermeasures to facial spoofing attacks boosted the development in face liveness detection [9].

Face image quality, facial motions, and scenic motions provide different liveness cues for face anti-spoofing. Various hand-engineered features have been proposed to describe liveness cues from the image quality-based or motion-based aspects, such as local binary pattern (LBP) and histogram of optical flow (HOOF) [10,11]. In this paper, shearlet transform is explored to discriminate between real faces and fake faces as an image quality descriptor. In contrast with the popularly used LBP, which is a local texture pattern feature, shearlet transform is more effective in representing distributed discontinuities, providing multi-scale and multi-directional anisotropic descriptions in images as directional wavelets [12]. On the aspect of motion cues, raw optical flow magnitude (OFM) is directly fed into an autoencoder neural network to learn motion-based liveness features in the cropped facial region and the whole scene, respectively. Compared with previous motion-based features, no motion hypothesis or statistic model is utilized to set motion priors in this paper. Therefore, the proposed motion-based liveness feature can achieve more generalization across face anti-spoofing databases with different scenic settings.

Due to varying attack scenarios and environment conditions, there is no absolutely superior face anti-spoofing technique. The combination of liveness features from the image quality-based and motion-based visual cues provides a promising direction to enhance the generalization and stability of a face anti-spoofing classifier. All the state-of-the-art face spoofing countermeasures

☆ This paper has been recommended for acceptance by Susanto Rahardja.

* Corresponding author.
E-mail address: lightedfeng@gmail.com (L. Feng).

http://dx.doi.org/10.1016/j.jvcir.2016.03.019
1047-3203/

take advantage of the feature fusion or score fusion approach, such as spatial–temporal texture descriptor [13], winning algorithms in the 2nd competition on countermeasures to 2D face spoofing attacks [9], and the complementary countermeasures [14]. Currently, a direct concatenation of feature vectors at feature level or a fusion at score level is wildly utilized. There are few attentions paid on developing a fusion strategy of different liveness features from multiple visual cues. In this paper, a feature fusion framework for integrating features from multiple categories of liveness cues is proposed using a neural network approach. The autoencoder neural network adopted is not only a supervised classifier but can also generate the bottleneck feature, which is a compressed sparse representation of the raw input for the neural network [15]. The bottleneck feature can represent raw inputs more effectively in a reduced dimension with a unit-scaled amplitude [16]. Hence, the liveness bottleneck features learned from different visual cues can be concatenated without scaling. The fused bottleneck features can be fed into a succeeding neural network to perform the final liveness detection. This hierarchical neural network can integrate liveness features from multiple visual cues and learn a complementary countermeasure to face spoofing attacks. Considering user-friendliness and convenience, the challenge-response and multimodality approaches are not discussed in this paper [17], because a face anti-spoofing method transparent to users is pursued here.

Compared to previous work, the contributions of this paper can be summarized as:

(1) Shearlet transform is utilized to perform face image quality assessment, providing a better image quality descriptor than the popularly used LBP.
(2) Motion-based liveness features are automatically learned using the neural network from raw optical flow information in the cropped facial region and the whole scene, respectively. With the pursuit of the generalization of the motion-based feature, no motion assumption or scenic model for face anti-spoofing is adopted.
(3) A feature fusion framework for integrating the image quality-based and the motion-based liveness cues is proposed using a hierarchical neural network. A higher face anti-spoofing classification accuracy is achieved by the proposed approach compared with the state-of-the-art methods.

The remainder of the paper is organized as follows. A brief literature review of the state-of-the-art face anti-spoofing methods is given in Section 2. The proposed feature fusion framework is explained in details in Section 3. Three public face anti-spoofing databases utilized in this paper are introduced in Section 4. Extensive experiments are conducted on the three public databases, and the corresponding results are reported in Section 5. Finally, a conclusion is drawn in Section 6.

2. Related work

Existing face anti-spoofing methods can be mainly classified into four categories: extra hardware-aided, image quality-based, motion-based, and multi-cues integration-based.

2.1. Extra hardware-aided

In addition to 2D images captured by regular cameras in the visible spectrum, extra hardware can provide other valuable information to distinguish between genuine faces and fake faces. Zhang et al. [18] selected two groups of light-emitting diodes (LED) with working spectra at 850 nm and 1450 nm as active light sources. Two corresponding photodiodes can detect the discriminant reflectance between real faces and fake materials. Lagorio et al. [19] captured 3D scanned points of face surfaces using an optoelectronic stereo system. 3D curvatures computed from acquired scanned data can distinguish between genuine faces and 2D spoofing mediums. Sun et al. [20] collected images in the visible spectrum and the infrared spectrum from a face simultaneously, with the combination of a visible camera and a thermal camera. Canonical correlation analysis of a pair of visible/thermal images using patched cross-modality was performed to detect face liveness. All the hardware-aided methods mentioned above obtained good performance on their private evaluation databases. However, the commercial applications of hardware-aided face anti-spoofing methods will be limited by requirements of setting up extra hardware.

2.2. Image quality-based

Spatial liveness information is extracted from static face images, with the expectation that fake face images displayed by spoofing mediums (paper, LCD screen, mask, etc.) will have image quality different from that of natural real face images, including sharpness, local artifacts, textureness, and statistical properties in some transform domains [21]. High frequency components of fake face images are much reduced in the Fourier spectrum compared with the cases of real face images [22]. Zhang et al. [7] transformed face images into a series of frequency bands using multiple difference of Gaussian (DoG) filters. Genuine and fake face images possess discriminant distributions in the DoG domain, and an overall equal error rate (EER) of 17% was achieved on the CASIA-FASD database. Micro-texture difference was observed between real and fake face images due to their surface differences [10]. LBP is a successful texture operator to describe micro-texture information. Block-based multi-scale LBP codes can obtain half total error rates (HTER) of 13.87%, 18.21%, and 0.95% on the REPLAY-ATTACK, CASIA-FASD, and 3D-MAD databases, respectively [6,8]. A parameterization combining 25 general image quality measures was proposed to perform the fake biometric detection, including face, iris, and fingerprint biometrics [21], and a comparable result with that of LBP can be achieved on the REPLAY-ATTACK database. Menotti et al. [23] extracted deep representations of raw face images through optimizing a deep convolutional neural network (CNN). This deep learning approach can gain a HTER of 0.75% on the REPLAY-ATTACK database

2.3. Motion-based

Motion patterns on the face or suspicious motion cues in the scenario are valuable temporal liveness information. Based on the region examined, motion-based face anti-spoofing can be grouped into two sub-categories, face motion-based and scene motion-based.

2.3.1. Face motion-based

Eye-blinking is a typical face motion cue wildly utilized in early work [24]. A human face is a non-rigid 3D object, exhibiting different optical flow trajectories compared with a 2D photo face. Kollreider et al. [25] assumed that the correlation between different facial components' motions is discriminant between real faces and photography faces. Bao et al. [26] proposed a motion model to describe the optical flow field of planar objects, and a divergence from this mode was assumed to exist in a real face's motion. Bharadwaj et al. [11] utilized Eulerian motion magnification to amplify subtle facial motions in a specialized frequency band. Macro- and micro-facial expressions presented by real faces can

be exaggerated and distinguished from distorted motion patterns on fake faces using the HOOF descriptor. A HTER of 1.25% was achieved on the REPLAY-ATTACK database.

2.3.2. Scene motion-based

The motion correlation between the user and background can indicate the presence of a spoofing attack. Kim et al. [27] supposed that the face and body region has low consistency with the background and the extracted background should not change in a preset authentication environment. Anjos et al. [28] utilized optical flow correlation between the user head and the background scene to detect photo face spoofing attacks, and a HTER of 1.49% was achieved on the PHOTO-ATTACK database. In practice, varying scenarios will be presented to face authentication systems, especially with mobile internet apps. One pre-defined face motion-based or scene motion-based model is not suitable for a wide variety of authentication environments.

2.4. Multi-cues integration-based

Liveness features relying on a single cue are not effective for every kind of face spoofing attack. The combination of complementary multi-cues from different aspects can solve several attack-specific sub-problems simultaneously. Hence the state-of-the-art results were achieved by the multi-cues integration-based approaches. Pereira et al. [13] utilized 3D-LBP based dynamic texture feature to describe the static facial texture and local facial motions at the same time. An EER of 10% and a HTER of 7.6% were achieved on the CASIA-FASD and REPLAY-ATTACK databases, respectively. Komulainen et al. [14] integrated the LBP-based texture feature and the motion correlation between the face and the background regions at score level. A HTER of 5.11% was obtained on the REPLAY-ATTACK database. In the 2nd competition on countermeasures to 2D face spoofing attacks [9], both the CASIA team and the LNMIIT team fused the LBP-based texture feature and motion features at feature level, and they both achieved perfect discrimination on the REPLAY-ATTACK database.

3. The proposed approach

The proposed multi-cues integration-based face anti-spoofing approach combines liveness features from three aspects: the shearlet-based image quality feature (SBIQF), the optical flow-based face motion feature, and the optical flow-based scene motion feature, as shown in the flow chart Fig.1, where $x_{k,i}$ represents the ith element in the input vector of the kth sub-network for a cue, $h_{k,i}^{(1)}$ is the learned ith primary feature activation in the hidden layer of the kth sub-network, $h_i^{(2)}$ is the learned ith primary feature activation in the second hidden layer of the integration neural network, and $P(y = C|x)$ is the probability of the C class (real/fake) with input x.

Firstly, a SBIQF vector is extracted from a normalized face image. Face coordinates are determined using Viola–Jones face detector and aligned with eyes-location [29]. A bottleneck representation for the SBIQF is obtained using the first sub-network in Fig. 1. Secondly, a face video is collected using the same face coordinates and normalization process as in the previous step. Dense optical flow is calculated between face frames with a fixed interval. An average OFM map describing the facial motion patterns is obtained with averaging the OFM information through the face video. And this average face OFM map is fed into the second sub-network to extract a bottleneck representation. Thirdly, an average scene OFM map is calculated from the scene video, which is the raw video where the face video is extracted. The scene OFM map is utilized as the input for the third sub-network to obtain a bottleneck representation. Lastly, all the three bottleneck representations from three different liveness cues are concatenated as a fused bottleneck feature, which is further fed into the subsequent neural network for liveness detection. The liveness status is finally determined using a two-class softmax classifier. As shown in Fig. 1, the three sub-networks before hidden layer II are locally connected with inputs from three different visual cues, respectively. And the three sub-networks are trained separately. At hidden layer II, the fused bottleneck feature is fully connected with the following network. The hidden layer II is trained layer-wisely with the hidden layer I. Details of core models in the flow chart are introduced as follows.

3.1. Autoencoder and softmax classifier

An autoencoder is a neural network trying to learn an approximation to the identity function, so as to output $\hat{x}$ that is similar to x [15], as shown in Fig. 2. The cost function for optimizing an autoencoder is given as

$$J_{sparse}(W,b) = J(W,b) + \beta \sum_{j=1}^{S_2} \mathrm{KL}(\rho \,\|\, \hat{\rho}_j) \tag{1}$$

where $J(W,b)$ is the cost function of an autoencoder for learning the identity function, $J_{sparse}(W,b)$ is the sparsity constrained cost function of an autoencoder, ρ is the sparsity parameter, $\hat{\rho}_j$ is the average activation of jth hidden unit, KL is the Kullback–Leibler divergence function for measuring the difference between ρ and $\hat{\rho}_j$, S_2 is the number of neurons in the hidden layer, and β is the weight of the sparsity penalty term. Through placing a dimension-reduced hidden layer and a sparsity constraint, a compressed sparse representation of the input can be obtained, as the bottleneck representation. Next, the learned bottleneck representation is fed as the input for a softmax classifier to build a classification neural network, as shown in Fig. 3.

A global fine-tuning of the overall neural network is performed by backpropagation using a labeled dataset. The autoencoder training can be treated as a pre-training process, which can provide a good initial solution for neural network optimization. Then the fine-tuning of parameters of the autoencoder and softmax classifier together using labeled data can further improve the bottleneck representation for liveness classification and reduce the training time. In this paper, both the three sub-networks for three visual cues and the feature integration network were trained using the way of pre-training followed by fine-tuning. Stacked autoencoders with multiple hidden layers can be trained layer-wisely. And the unfolded stacked autoencoders can be fine-tuned together with a softmax classifier using labeled data. The layer-wise pre-training followed by fine-tuning is the core idea in deep learning [30]. Due to the sigmoid activity function used in neural networks, the bottleneck representation is automatically scaled between 0 and 1, which is suitable for fusing multi-cues features with different scales.

3.2. Shearlet-based image quality feature

Curvilinear singularities exhibited in images cannot be sparsely approximated using traditional wavelet, due to a lack of direction descriptor. To overcome drawbacks of wavelet, shearlet has been proposed in recent years, with an ability of efficiently capturing anisotropic features in multidimensional data [12]. Shearlet transform has been successfully utilized to perform non-reference image quality assessment as a state-of-the-art method [31]. Compared with real faces, spoofing faces may have sharpness reduction, textureness differences, additive noises, and artifacts due to face reproducing on spoofing mediums. Compared with LBP and DoG, shearlet can better describe curvilinear singularities,

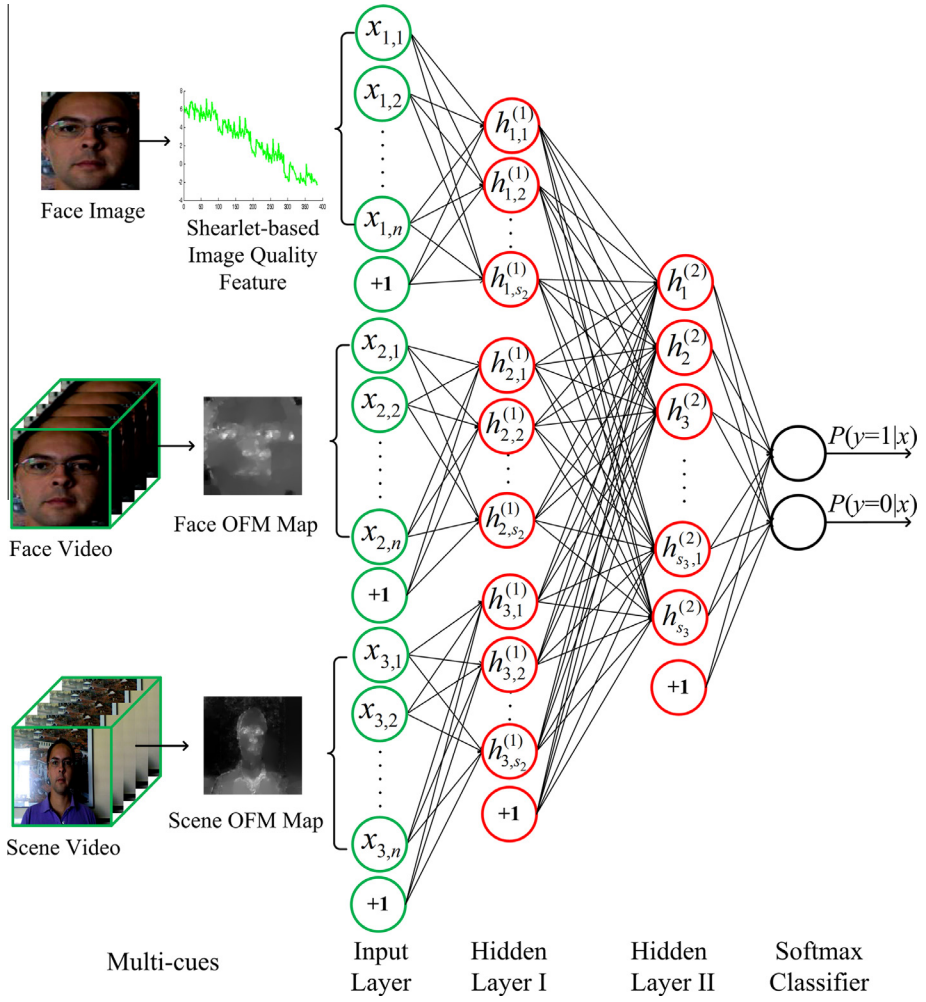

Fig. 1. The flow chart of the multi-cues integration-based face anti-spoofing method using neural networks.

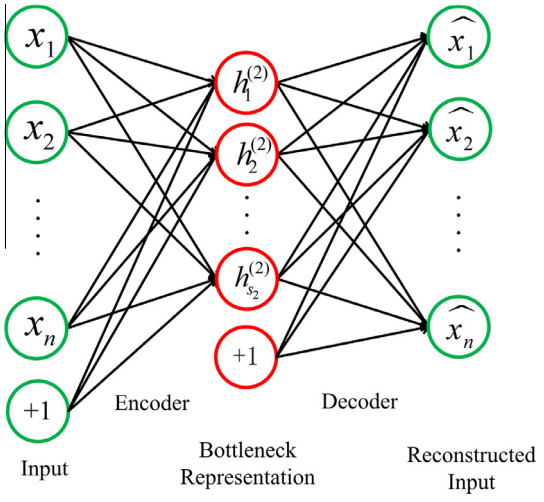

Fig. 2. Autoencdoer architecture.

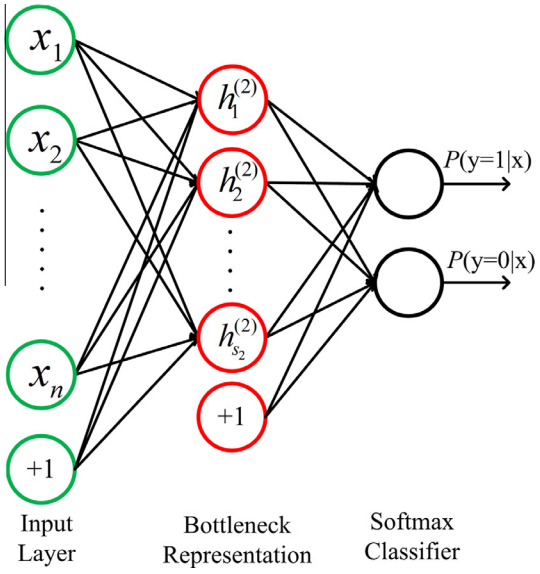

Fig. 3. Fine-tuning of autoencoder neural network with softmax classifier.

including edges, textures, and artifacts. Image quality assessment based on shearlet has an advantage in detecting blurred edges and distorted textures on spoofing faces, anisotropic artifacts caused by face reproducing (video coding, printing, etc.), textureness differences between real facial skin and spoofing mediums, etc. At the same time, shearlet can also describe isotropic noises and artifacts on spoofing faces. Hence the proposed image quality-based liveness feature is based on shearlet. For a

two-dimensional image, the affine systems with composite dilations are the collections of the form [32]:

$$SH_\phi f(a,s,t) = \langle f, \phi_{a,s,t} \rangle, \quad a > 0,\ s \in R,\ t \in R^2 \tag{2}$$

where f is the image and the analyzing factor $\phi_{a,s,t}$ is the shearlet basis, which is defined as

$$\phi_{a,s,t}(x) = |\det M_{a,s}|^{-\frac{1}{2}} \phi(M_{a,s}^{-1} x - t) \tag{3}$$

where $M_{a,s} = B_s A_a = \begin{pmatrix} a & \sqrt{a}s \\ 0 & \sqrt{a} \end{pmatrix}$ and $A_a = \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix}$, $B_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$. A_a is the anisotropic dilation matrix and B_s is the shear matrix. Since the analyzing functions associated with the shearlet transform are anisotropic and defined at different scales, locations, and orientations, shearlet is able to detect directional information and account for geometry of multidimensional functions.

Starting with the shearlet transform of a face image into different sub-bands, the calculation process of the SBIQF is summarized in Fig. 4. Each element in a red[1] box in a sub-band is defined as

$$x(a,s,b) = \frac{\sum |SH_\phi f(a,s,b)|}{m^2} \tag{4}$$

where $a = 1,\ldots,A$ is the scale index (excluding coarsest scale), $s = 1,\ldots,S$ is the direction index and $b = 1,\ldots,(M/m)^2$ is the block index of each sub-band. M is the size of the square image and m is the size of each red block. $SH_\phi f(a,s,b)$ are the shearlet coefficients of each red block. Mean pooling of shearlet coefficients is performed in each red block. The pooled values are concatenated as a vector subjected to a logarithmic nonlinearity:

$$\text{SBIQF} = \log_2(x_1, x_2 \ldots, x_N) \tag{5}$$

where $N = A \times S \times (M/m)^2$ is the total number of red blocks.

3.3. Optical flow-based motion feature

Optical flow can describe local image motions based on local derivatives in a given image sequences. With the assumption of brightness constancy and spatial smoothness, optical flow can be calculated through solving motion constraint equation:

$$\frac{\partial I}{\partial x} v_x + \frac{\partial I}{\partial y} v_y + \frac{\partial I}{\partial t} = 0 \tag{6}$$

where I is the image intensity at (x,y,t), v_x and v_y are the x and y components of optical flow, describing a local pixel translation. A dense optical flow technique based on iterative reweighted least squares (IRLS) method is adopted in this paper [33]. Optical flow describes the motion direction and motion magnitude simultaneously. Only motion magnitude information is employed here.

Compared with previous hand-crafted motion-based face liveness features, the proposed optical flow-based motion feature does not depend on any pre-defined model or prior assumption, such as eye-blinking on real faces [24], non-rigid facial expression [25], planar motion patterns on a photography face [26], and low motion consistency between a real face and the background [27], because it is hard to build a universal motion model to describe motion cues for face anti-spoofing in varying spoofing scenarios. The OFM map can capture every motion on the face or the scene. A neural network is good at learning implicit patterns, which is able to recognize motion cues for face liveness detection with proper training. Hence the OFM map is selected to describe the motion-based cues for learning to deal with different spoofing attacks.

[1] For interpretation of color in Figs. 4 and 10, the reader is referred to the web version of this article.

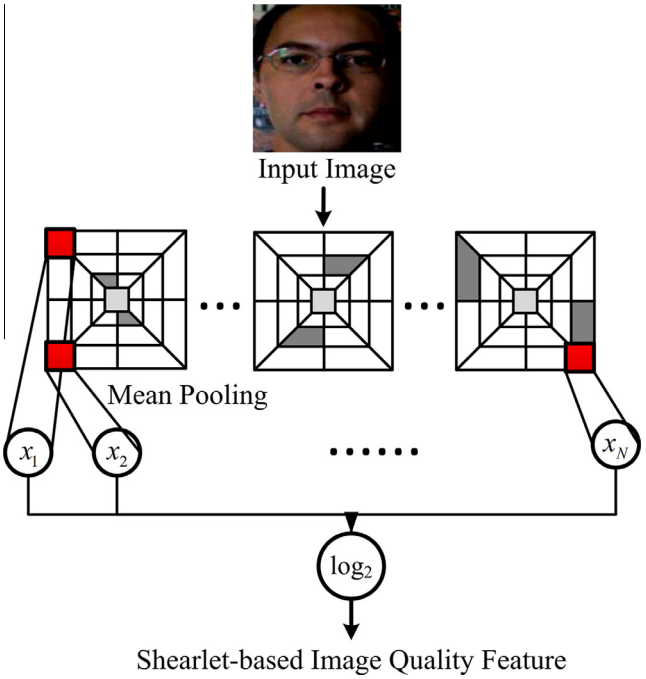

Fig. 4. Calculation process of shearlet-based image quality feature.

A short video ($\leqslant$2 s) is recorded during liveness detection. After that, a face video is extracted using a simple eyes-location normalization. Dense optical flow calculation is performed between two frames with a fixed interval, and an OFM map can be obtained to describe pixel-level motion trajectories between the two frames. Several face OFM maps are generated with pairs of frames in the face video using dense optical flow. An average of these face OFM maps is utilized to record motion patterns on the face, as shown in Fig. 5(a). An average scene OFM map is calculated from the whole-frame video, as shown in Fig. 5(b). Scenic motion cues in the background can be recorded in the average scene OFM map. Face region or body region are not excluded for calculating scene OFM maps, because it is difficult to define a uniform foreground/background model in face anti-spoofing. The average face/scene OFM map are columnized into two input vectors for face motion-based and scene motion-based sub neural networks, respectively.

Based on how the fake face is represented with the spoofing medium, 2D face spoofing attacks can be classified into two categories: close-up and scenic spoofs. A close-up spoof describes only the facial area which is presented to the sensor, during which photography/screen edges and the attacker's hands are visible on the scene. In contrast with close-up spoofs, scenic face spoofs incorporate the background scene in the face spoof. A resulting scenic fake face is placed near to the sensor to hide medium boundaries or human hands.

Some typical motion patterns will appear in face spoofing attacks, which are distinguishable from motion patterns in real accesses. For example, Fig. 6(a) shows an average OFM map on a real face, where non-rigid local facial motions concentrate on facial components. A face displayed on a photographic paper has a global movement due to involuntary hand-shaking. Then comparable OFM can be observed on both the face and the background neighboring the face, as shown in Fig. 6(b). Specular reflection often occurs on mirror-like screens of video-replay devices, which can also be detected using the OFM map as shown in Fig. 6(c). A face mask will cover local facial motions except eyes movements, and a uniform motion pattern on the mask can be recorded on the

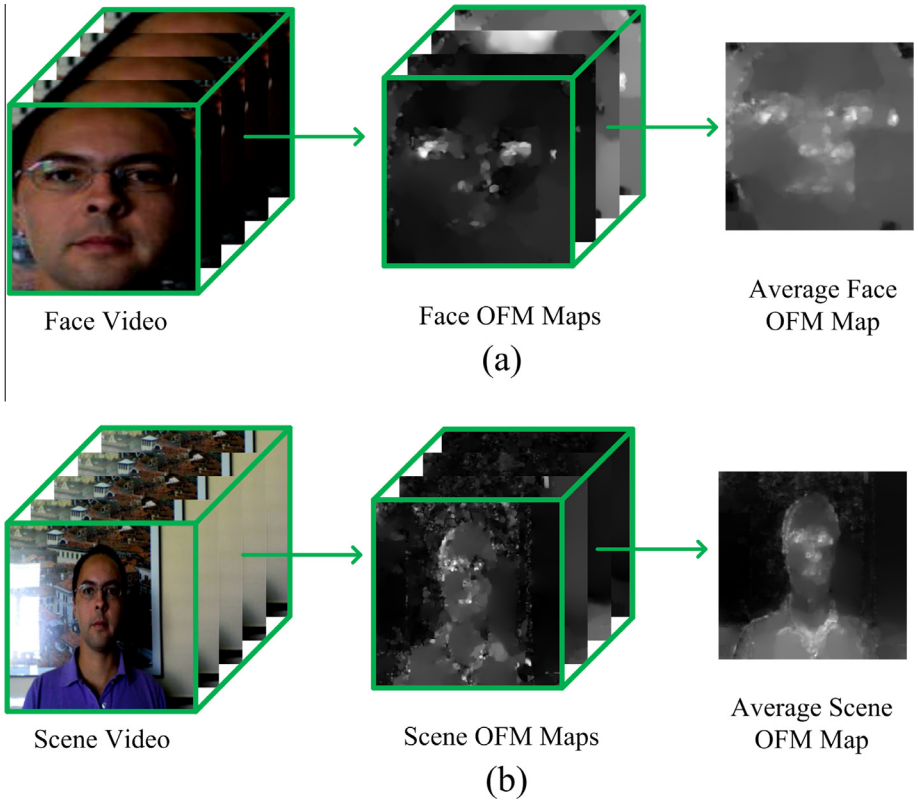

Fig. 5. The calculation of an average OFM map: (a) on the face; (b) on the scene.

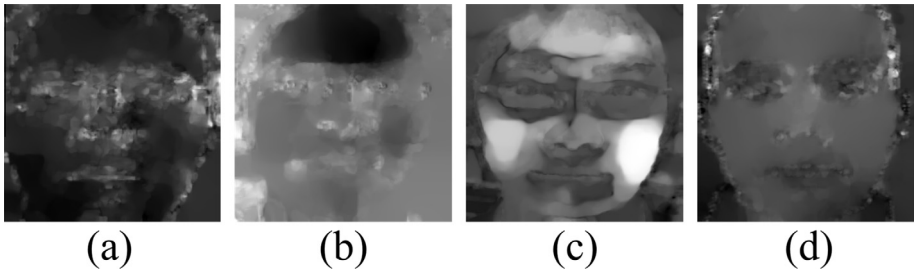

Fig. 6. Typical facial motion patterns on a real face and three spoofing faces: (a) a real face; (b) a hand-held photography face; (c) a video-replay face displayed on a high-definition screen; (d) a masked face.

OFM map, as shown in Fig. 6(d). A clear face profile appears on the scene OFM map for a real access, as shown in Fig. 7(a). For a scenic spoof, if the display medium is not supported stably, a global movement across the frame can be detected as shown in Fig. 7 (b). For a close-up spoof, suspicious movements of hands and borders of a video-replay device can be observed on the scene OFM map, as shown in Fig. 7(c).

4. Face anti-spoofing databases

In order to evaluate performance of the proposed face anti-spoofing approach, three public face anti-spoofing databases are utilized as benchmarks, including the REPLAY-ATTACK database, the CASIA-FASD database, and the 3D-MAD database.

4.1. REPLAY-ATTACK database

This database collected 1200 short videos of both real accesses and face spoofing attacks from 50 subjects, recorded using a webcam [6]. The videos were recorded in two different lighting conditions: (1) Controlled, with a uniform background and artificial illumination; (2) Adverse, with a complex background and natural illumination. Three different kinds of attacks with two different support conditions were considered. Three kinds of attacks include: (1) Print: high-resolution photographs printed on A4 papers were presented to the camera; (2) Mobile: photos and videos taken and displayed using a smartphone were presented to the camera; (3) Highdef: high-resolution photos and videos displayed using a tablet were presented to the camera. Two support conditions include: (1) Hand-held: the attack media was held by hands; (2) Fixed: the attack media was attached on a fixed support. Some example frames from real accesses and spoofing attacks in different scenarios are shown in Fig. 8.

All subjects in REPLAY-ATTACK database are partitioned into three non-overlapped sub-sets with 15, 15, and 20 subjects respectively: (1) Train, to tune parameters of the classifier; (2) Development, to fix the decision threshold; (3) Test, to evaluate final classification performance.

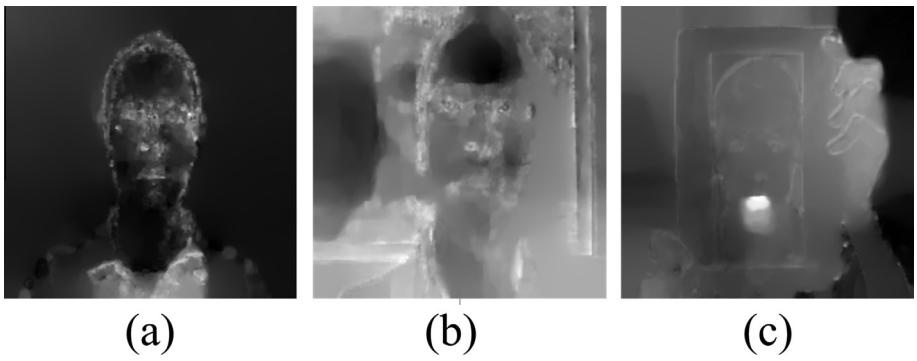

Fig. 7. Typical scenic motion patterns for a real access and two spoofing attacks: (a) a real person; (b) a scenic hand-held photography; (c) a close-up hand-held video-replay device.

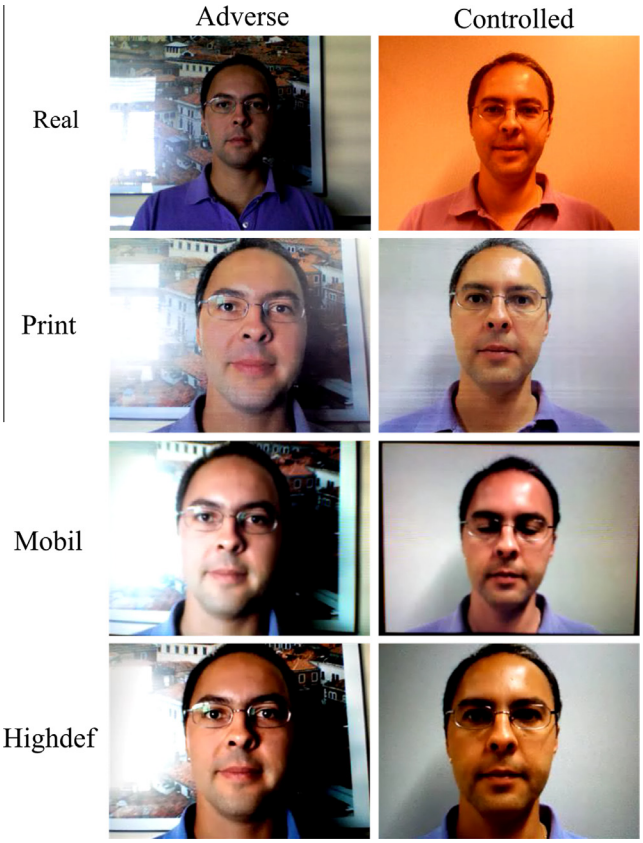

Fig. 8. Examples of real accesses and spoofing attempts in different scenarios in the REPLAY-ATTACK database.

4.2. CASIA-FASD database

This database consists of 600 short videos from 50 clients [7]. Their real accesses and corresponding three different kinds of spoofing attempts were recorded using three digital cameras with different imaging quality. The three kinds of attacks include: (1) Warped-photo: face photographs printed on copper papers were presented to the camera, simulating facial motions through warping photos; (2) Cut-photo: the eyes region was cut off from the photography for exhibiting eye-blinking; (3) Video: high-quality genuine videos were displayed using a high-resolution tablet presented to the camera. Three different imaging quality conditions were implemented using an imaging device of (1) Low-quality, (2) Middle-quality, and (3) High-quality, respectively. Example frames from genuine and fake videos in different scenarios are shown in Fig. 9. The CASIA-FASD database is divided into (1) Train sub-set and (2) Test sub-set with 20 and 30 independent subjects, respectively.

4.3. 3D-MAD database

This database is composed of 255 short videos of real accesses and mask attacks from 17 subjects [8]. Color images and depth maps were captured simultaneously using a natural user interface device. Only color images are utilized in this paper. Example frames of a real face and a masked face are shown in Fig. 10. The 3D-MAD database is randomly partitioned into three non-overlapped sub-sets with 7, 5, and 5 subjects respectively: (1) Train, to optimize the classifier; (2) Development, to fix the decision threshold; (3) Test, to report the spoofing detection result.

4.4. Protocol

HTER is advised to objectively evaluate a proposed countermeasure on the REPLAY-ATTACK and 3D-MAD databases. HTER is defined as

$$\mathrm{HTER} = \frac{\mathrm{FAR}(\tau, D) + \mathrm{FRR}(\tau, D)}{2} \tag{7}$$

where FAR means the false acceptance rate, FRR means the false rejection rate, D is the test sub-set, and τ is the decision threshold. The value of τ is determined on the EER using the development sub-set. The EER is the HTER subjected to that the FAR equals the FRR. As mentioned above, there is no explicit partition for the train, development, and test sub-sets in the 3D-MAD database. Hence, the results reported on the 3D-MAD database are the average HTER over 10 iterations of a random database partition. There is a lack of a development sub-set in the CASIA-FASD database. Thereby, the EER is utilized to report the evaluation results on the CASIA-FASD database instead of the HTER, as advised in its protocol.

5. Experimental results

The experiments have been designed with a two-fold objective. First, face anti-spoofing performance of the three proposed face liveness features were evaluated, including the SBIQF, the average face OFM map, and the average scene OFM map. Then the proposed multi-cues integration-based face anti-spoofing approach was evaluated and compared with the state-of-the-art face anti-spoofing algorithms. Second, the multi-cues fusion strategy was discussed through the comparisons among the raw feature fusion, the score fusion, and the proposed bottleneck feature fusion.

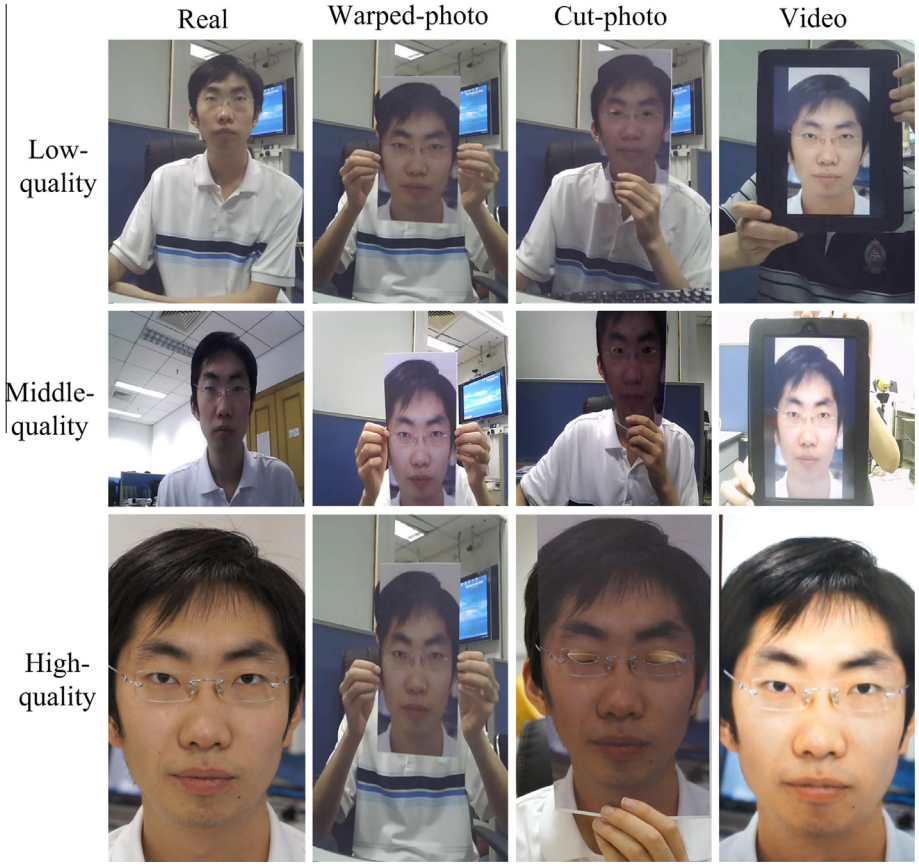

Fig. 9. Examples of real accesses and spoofing attempts in different scenarios in the CASIA-FASD database.

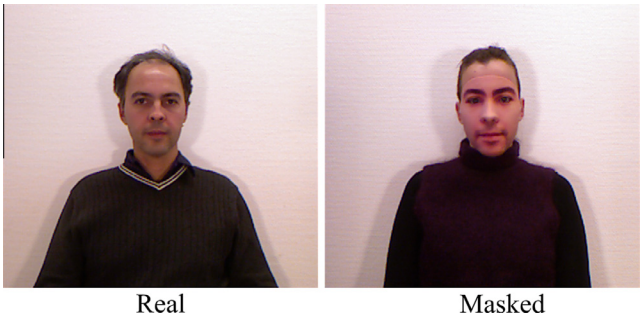

Fig. 10. Examples of a real access and a mask attack in the 3D-MAD database.

5.1. Setting of experimental parameters

Consecutive 60 frames were randomly selected from each video in databases to generate a scene (whole-frame) video. The face coordinates determined in the initial frame using Viola–Jones algorithm with eyes-location alignment was utilized for all the 60 frames to extract a face video from the scene video. An OFM map is calculated between two frames. For calculating an average face/scene OFM map, 6 face/scene frames were firstly selected from 30 consecutive face/scene frames with a fixed interval of 5 frames. Then 5 face/scene OFM maps were calculated consecutively using the 6 selected face/scene frames. An average face/scene OFM map was obtained through averaging these 5 face/scene OFM maps. And a SBIQF was extracted from the first one of the 6 selected face frames. Thus, a SBIQF, an average face OFM map, and an average scene OFM map were extracted from consecutive 30 frames as an input for the proposed multi-cues integration neural network. With 10 times of slides of 3 frames within the total 60 frames, 10 sets of inputs were thus obtained. And the final liveness status in a video was determined by averaging the liveness detection scores over 10 sets of inputs. For single performance evaluation of each proposed liveness feature, the final video-score was also averaged over 10 inputs. For the SBIQF extraction, gray-scale face frames were normalized into 256×256 pixels. The scale number A, the direction number S, and the pooling block-size are set as 4, 6, and 64×64 pixels respectively, resulting in the length of the SBIQF as 384. For the average face/scene OFM map calculation, face/scene videos were down-sampled into a resolution of 32×32 pixels to reduce computational cost. Hence, the input vector length of the columnized average face/scene OFM map is 1024.

For building the multi-cues integration neural network (Fig. 1), three sub-networks with one hidden layer of 60 neurons were trained for three visual cues, respectively. After that, the second hidden layer with 80 neurons was trained with the fused bottleneck representations from multi-cues as inputs. In this paper, all neural networks were trained by the means of pre-training using the autoencoder followed by fine-tuning through the whole classification network using labeled data. All inputs for neural networks were normalized to zero-means and unit-variances. The weight decay parameter for autoencoders and softmax classifiers is 3e−5. The sparsity parameter is 0.1 and the weight of sparsity penalty term is 3. Support vector machine (SVM) was utilized for a performance comparison. The parameters of SVM kernels were set using grid-search [34].

5.2. Face anti-spoofing performance evaluation

The effectiveness of the proposed three face liveness features for three different visual cues were tested separately. Then

Table 1
Performance of various face anti-spoofing approaches (%).

Approach	REPALY-ATTACK		CASIA-FASD	3D-MAD	
	Devel. (EER)	Test (HTER)	Test (EER)	Devel. (EER)	Test (HTER)
LBP + SVM [6,35][a]	8.55	11.75	18.50	–	0.95
DOG + SVM [7][a]	–	–	17.00	–	–
LBP-TOP + SVM [13][a]	7.88	7.60	10.00	–	–
Complementary Countermeasures + LDA [14][a]	4.57	5.11	–	–	–
Component Dependent Descriptor + SVM [35][a]	–	–	11.8	–	–
Motion Magnification + LDA [11][a]	0	1.25	–	–	–
Deep Representation + CNN [24][a]	–	0.75	–	–	0
Fusion of Texture and Motion + SVM [9][a]	0	0	–	–	–
SBIQF + SVM	3.17	7.13	17.78	0	0.50
SBIQF + NN	3.83	6.13	15.5	0	0
Face OFM Map + NN	3.83	2.50	19.81	7.00	7.00
Scene OFM Map + NN	3.50	6.16	18.33	3.00	4.00
Multi-cues integration + NN	0.83	0	5.83	0	0

[a] Results as reported in citation (under the same protocol). Each approach is presented as Feature Description + Classifier. LDA – Linear Discriminant Analysis. CNN – Convolutional Neural Network. NN – Neural Network.

Table 2
Performance of different multi-cues fusion strategies for face anti-spoofing (%). (NN – Neural Network).

Fusion strategy	REPALY-ATTACK		CASIA-FASD	3D-MAD	
	Devel. (EER)	Test (HTER)	Test (EER)	Devel. (EER)	Test (HTER)
Raw Feature Fusion + SVM	4.33	8.33	9.07	0	2.00
Raw Feature Fusion + NN	1.50	2.75	10.93	0	2.00
Score Fusion + NN	3.00	2.50	10.93	0	0
Bottleneck Feature Fusion + NN	0.83	0	5.83	0	0

performance of the proposed multi-cues integration-based face anti-spoofing approach was evaluated on the three databases. All these results are given in Table 1. For a performance comparison, the results of the baseline algorithms in databases and the state-of-the-art countermeasures to face spoofing attacks are also listed in Table 1.

In order to make a comparison between LBP and SBIQF, both of them utilized SVM to perform the liveness classification. And the proposed SBIQF has a better ability in describing the image quality discrepancy between genuine faces and fake faces, showing lower HTER/EER on all three databases compared with LBP. When SVM was replaced with a neural network, performance of SBIQF obtained some improvement.

On the REPLAY-ATTACK database, a perfect classification has been achieved through fusing texture-based features and motion-based features, proposed by two teams in the 2nd competition on countermeasures to 2D face spoofing attacks. The proposed multi-cues integration-based approach can also achieve a HTER of 0% on the test sub-set of the REPLAY-ATTACK database. On the CASIA-FASD database, best performance in previous work was achieved by the LBPs from three orthogonal planes (LBP-TOP) method, exploring the spatial and temporal LBP distributions simultaneously. The proposed multi-cues integration-based approach achieved an EER of 5.83%, which is better than the LBP-TOP method. On the 3D-MAD database, a perfect discrimination was achieved by the deep representation method, which utilized a deep CNN to learn local texture-based features from raw face images. Both the proposed SBIQF and the proposed multi-cues integration neural network can obtain a HTER of 0% on the 3D-MAD database. This implies that image quality-based features work effectively on distinguishing between real faces and face masks.

On both the REPLAY-ATTACK database and the CASIA-FASD database, the proposed multi-cues integration-based approach achieved a huge performance improvement in liveness detection compared with all the three input liveness features. This result illustrates the effectiveness of the proposed multi-cues integration approach using a hierarchical neural network and the complementarity of the SBIQF, the average face OFM map, and the average scene OFM map.

5.3. Multi-cues fusion strategy

A bottleneck feature fusion was performed in the proposed multi-cues integration neural network. In order to investigate the effectiveness of the bottleneck feature fusion, a variety of multi-cues fusion strategies for face anti-spoofing were evaluated for a comparison, as shown in Table 2.

Raw feature fusion means that the proposed three liveness features are directly concatenated on feature level without learning bottleneck representations. Then the concatenated raw features were fed into SVM or neural networks for classification. Score fusion means that the proposed three liveness features are fed into three separate neural networks for the face anti-spoofing classification. Then the scores from the three neural networks are fused using logistic regression. Since the proposed multi-cues integration neural network has two hidden layers, stacked autoencoders with two hidden layers of [60 60] neurons were utilized for raw feature fusion and score fusion for a fair comparison. In Table 2, the proposed bottleneck feature fusion achieved best performance on all the three databases. This illustrates the sparsity, reduced-dimension, and unit-scale of bottleneck representations are helpful in feature fusion. The score fusion strategy obtained comparable or even better results than raw feature fusion strategies, which implies that a proper feature fusion strategy is critical for multi-cues integration for face anti-spoofing.

6. Conclusion and future directions

With the rapid development of face anti-spoofing techniques, the threats of spoofing attacks will also increase in the diversity,

the reality, and the sophistication. It will be hard to select one technique over the others. Hence, the combination of several complementary countermeasures is a promising approach. In this paper, an effective multi-cues integration neural network is proposed for face anti-spoofing, which fuses the SBIQF, the average face OFM map, and the average scene OFM map using bottleneck representations. Extensive experiments were performed on three publicly available databases to evaluate the proposed multi-cues integration-based face anti-spoofing approach. A perfect discrimination between real accesses and spoofing attacks were achieved on the REPALY-ATTACK database and the 3D-MAD database. An EER of 5.83% was achieved on the CASIA-FASD database, which is better than the state-of-the-art methods. Compared with LBP, the proposed SBIQF is better in describing the image quality discrepancy between real faces and fake faces. The bottleneck feature fusion strategy obtained better performance than the strategies of raw feature fusion and score fusion. The proposed face anti-spoofing approach has been implemented using a combination of C programs and MATLAB programs on a desktop computer with a 2.67 GHz processor and 16 GB of memory. A demo video is available at https://youtu.be/151USnKDKZY. The proposed face anti-spoofing approach can be easily combined with a face identification module via sparse representation [36], as shown in the demo face authentication system. The face identify, user existence, and liveness status can be determined within 4.15 s, including a 2-s-long video collection process with 30 frames per second. In future work, other advanced neural networks will be investigated to improve face anti-spoofing performance, such as the convolutional neural network and the long short-term memory (LSTM) network [37,38], which may be more effective in learning face liveness features.

Acknowledgements

We would like to sincerely thank Idiap Research Institute and CBSR of Chinese Academy of Sciences for the share of their face anti-spoofing databases. Many thanks for the helpful comments from all reviewers.

Reference

[1] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification–verification, in: Proc. Annual Conference on Neural Information Processing Systems, Montreal, Canada, 2014, pp. 1988–1996.
[2] R. Ding, D.K. Du, Z. Huang, Z. Li, K. Shang, Variational feature representation-based classification for face recognition with single sample per person, J. Vis. Commun. Image Represent. 30 (2015) 35–45.
[3] M. De Marsico, C. Galdi, M. Nappi, D. Riccio, FIRME: face and iris recognition for mobile engagement, Image Vis. Comput. 32 (12) (2014) 1161–1172.
[4] N.M. Duc, B.Q. Minh, Your face is NOT your password, in: Proc. Conference on Black Hat, Las Vegas, USA, 2009, pp. 1–16.
[5] J. Galbally, S. Marcel, J. Fierrez, Biometric antispoofing methods: a survey in face recognition, IEEE Access 2 (2014) 1530–1552.
[6] I. Chingovska, A. Anjos, S. Marcel, On the effectiveness of local binary patterns in face anti-spoofing, in: Proc. International Conference of the Biometrics Special Interest Group, Darmstadt, Germany, 2012, pp. 1–7.
[7] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, S.Z. Li, A face anti-spoofing database with diverse attacks, in: Proc. International Conference on Biometrics, New Delhi, India, 2012, pp. 26–31.
[8] N. Erdogmus, S. Marcel, Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect, in: Proc. IEEE International Conference on Biometrics: Theory, Applications and Systems, Arlington, USA, 2013, pp. 1–6.
[9] I. Chingovska, The 2nd competition on counter measures to 2D face spoofing attacks, in: Proc. International Conference on Biometrics, Madrid, Spain, 2013, pp. 1–6.
[10] J. Maatta, A. Hadid, M. Pietikainen, Face spoofing detection from single images using micro-texture analysis, in: Proc. International Joint Conference on Biometrics, Washington DC, USA, 2011, pp. 1–7.
[11] S. Bharadwaj, T.I. Dhamecha, M. Vatsa, R. Singh, Computationally efficient face spoofing detection with motion magnification, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013, pp. 105–110.
[12] G. Easley, D. Labate, W.Q. Lim, Sparse directional image representations using the discrete shearlet transform, Appl. Comput. Harmon. Anal. 25 (1) (2008) 25–46.
[13] T.F. Pereira, J. Komulainen, A. Anjos, J.M.D. Martino, A. Hadid, M. Pietikainen, S. Marcel, Face liveness detection using dynamic texture, EURASIP J. Image Video Process. 1 (2014) 2014.
[14] J. Komulainen, A. Hadid, M. Pietikainen, A. Anjos, S. Marcel, Complementary countermeasures for detecting scenic face spoofing attacks, in: Proc. International Conference on Biometrics, Madrid, Spain, 2013, pp. 1–7.
[15] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[16] D. Yu, M.L. Seltzer, Improved bottleneck features using pretrained deep neural networks, in: Proc. Annual Conference of the International Speech Communication Association, Florence, Italy, 2011, pp. 237–240.
[17] D.J. Kim, K.W. Chung, K.S. Hong, Person authentication using face, teeth and voice modalities for mobile device security, IEEE Trans. Consumer Electron. 56 (4) (2010) 2678–2685.
[18] Z. Zhang, D. Yi, Z. Lei, S.Z. Li, Face liveness detection by learning multispectral reflectance distributions, in: Proc. IEEE International Conference on Automatic Face and Gesture Recognition, Santa Barbara, USA, 2011, pp. 436–441.
[19] A. Lagorio, M. Tistarelli, M. Cadoni, C. Fookes, S. Sridharan, Liveness detection based on 3D face shape analysis, in: Proc. International Workshop on Biometrics and Forensics, Lisboa, Portugal, 2013, pp. 1–4.
[20] L. Sun, W. Huang, M. Wu, TIR/VIS correlation for liveness detection in face recognition, in: Proc. International Conference on Computer Analysis of Images and Patterns, Seville, Spain, 2011, pp. 114–121.
[21] J. Galbally, S. Marcel, J. Fierrez, Image quality assessment for fake biometric detection: application to iris, fingerprint, and face recognition, IEEE Trans. Image Process. 23 (2) (2014) 710–724.
[22] J. Li, Y. Wang, T. Tan, A.K. Jain, Live face detection based on the analysis of Fourier spectra, Proc. SPIE 5404 (2004) 296–303.
[23] D. Menotti, G. Chiachia, A. Pinto, W.R. Schwartz, H. Pedrini, A.X. Falcao, A. Rocha, Deep representations for iris, face, and fingerprint spoofing detection, IEEE Trans. Inform. Forensics Sec. 10 (4) (2015) 864–879.
[24] G. Pan, L. Sun, Z. Wu, S. Lao, Eyeblink-based anti-spoofing in face recognition from a generic webcamera, in: Proc. International Conference on Computer Vision, Rio de Janeiro, Brazil, 2007, pp. 1–8.
[25] K. Kollreider, H. Fronthaler, J. Bigun, Non-intrusive liveness detection by face images, Image Vis. Comput. 27 (3) (2009) 233–244.
[26] W. Bao, H. Li, N. Li, W. Jiang, A liveness detection method for face recognition based on optical flow field, in: Proc. International Conference on Image Analysis and Signal Processing, Taizhou, China, 2009, pp. 233–236.
[27] Y. Kim, J.H. Yoo, K. Choi, A motion and similarity-based fake detection method for biometric face recognition systems, IEEE Trans. Consumer Electron. 57 (2) (2011) 756–762.
[28] A. Anjos, M.M. Chakka, S. Marcel, Motion-based counter-measures to photo attacks in face recognition, IET Biomet. 3 (3) (2014) 147–158.
[29] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, USA, 2001, pp. I-511–I-518.
[30] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks, in: Proc. Annual Conference on Neural Information Processing Systems, Vancouver, Canada, 2007, pp. 153–160.
[31] Y. Li, L.M. Po, X. Xu, L. Feng, No-reference image quality assessment using statistical characterization in the shearlet domain, Signal Process.: Image Commun. 29 (7) (2014) 748–759.
[32] G. Kutyniok, D. Labate, Shearlets, first ed., Birkhäuser-Verlag, Boston, 2012.
[33] C. Liu, Beyond Pixels: Exploring New Representations and Applications for Motion Analysis (Doctoral thesis), Massachusetts Institute of Technology, 2009.
[34] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27-1–27-27.
[35] J. Yang, Z. Lei, S. Liao, S.Z. Li, Face liveness detection with component dependent descriptor, in: Proc. International Conference on Biometrics, Madrid, Spain, 2013, pp. 1–6.
[36] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2008) 210–227.
[37] G. Chen, C. Parada, T.N. Sainath, Query-by-example keyword spotting using long short-term memory networks, in: Proc. International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, 2015, pp. 5236–5240.
[38] Y. Taigman, M. Yang, M.A. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014, pp. 1701–1708.