Journal of Electronic Imaging

JElectronicImaging.org

Light field-based face liveness detection with convolutional neural networks

Mengyang Liu
Hong Fu
Ying Wei
Yasar Abbas Ur Rehman
Lai-man Po
Wai Lun Lo

SPIE. IS&T imaging.org

Mengyang Liu, Hong Fu, Ying Wei, Yasar Abbas Ur Rehman, Lai-man Po, Wai Lun Lo, "Light field-based face liveness detection with convolutional neural networks," *J. Electron. Imaging* **28**(1), 013003 (2019), doi: 10.1117/1.JEI.28.1.013003.

Light field-based face liveness detection with convolutional neural networks

Mengyang Liu,[a,b] Hong Fu,[a,*] Ying Wei,[c,d] Yasar Abbas Ur Rehman,[b] Lai-man Po,[b] and Wai Lun Lo[a]
[a]Chu Hai College of Higher Education, Department of Computer Science, Hong Kong SAR, China
[b]City University of Hong Kong, Department of Electronic Engineering, Hong Kong SAR, China
[c]Shandong University, School of Information Science and Engineering, Jinan, China
[d]Center of International Joint Research on Intelligent Perception and Information Processing, Jinan, China

Abstract. Face recognition based-access systems have been used widely in security systems as the recognition accuracy can be quite high. However, these systems suffer from low robustness to spoofing attacks. To achieve a reliable security system, a well-defined face liveness detection technique is crucial. We present an approach for this problem by combining data of the light-field camera (LFC) and the convolutional neural networks in the detection process. The LFC can detect the depth of an object by a single shot, from which we derive meaningful features to distinguish the spoofing attack from the real face, through a single shot. We propose two features for liveness detection: the ray difference images and the microlens images. Experimental results based on a self-built light-field imaging database for three types of the spoofing attacks are presented. The experimental results show that the proposed system gives a lower average classification error (0.028) as compared with the method of using hand-crafted features and conventional imaging systems. In addition, the proposed system can be used to classify the type of the spoofing attack. [DOI: 10.1117/1.JEI.28.1.013003]

Keywords: face liveness detection; light filed camera; convolutional neural networks; face spoofing attack.

Paper 180467 received Jun. 21, 2018; accepted for publication Dec. 5, 2018; published online Jan. 8, 2019.

1 Introduction

Face recognition is one of the most common biometric methods to identify or to verify individuals because of its noninvasive property. Due to the development of the advanced image sensors and the sophisticated image processing techniques, acquiring facial information becomes easy and this approach can eliminate the risk of forgetting login key and password information for a computer user account. For these reasons, using facial information to access security systems becomes popular, and this approach can make the systems more convenient and reliable.[1]

Although face recognition systems are widely used in security systems such as intelligent entrance guard systems at companies and schools, bank user account logins, and website account login processes, etc., they are still vulnerable to various intentional attacks. For the system, it is difficult to detect whether the face in front of the camera is a bona fide face or a spoofing attack. In practice, even some simple methods can pass the security system. There are four common types of flat face spoofing attacks: (i) using a printed photo, (ii) displaying a photo by using high-definition (HD) screen, (iii) warped print attacks, and (iv) displaying a video using an HD screen. A more challenging type of attack, which is the three-dimensional (3-D) facial mask attack,[2] has been raised recently. All these attacks can provide facial information from which the system can get a valid recognition result. For this reason, a technical approach for the defense against spoofing attacks is necessary for a face recognition security system. Liveness detection aims to distinguish the testing face of a bona fide person from a spoofing attack.

Past research in this area can be classified into three approaches—the systems with extra devices, 2-D information methods, and 3-D information methods. Extra device approaches have been used in industries, which use infrared sensors[3] or an extra camera.[4] Two-dimensional (2-D) image-based methods have relatively low computational cost and can be embedded in portable devices. The 2-D image-based methods can be further separated into three main categories[5] based on the types of liveness indicator they used: motion analysis,[6] texture analysis,[7] and life sign detection. As the definition and color range of screens becomes higher and wider, it is more difficult to detect the texture difference between a real object and a screen image. In addition, for the motion analysis and life sign detection methods, the system cannot perform well when the spoofing attack is a video sequence. In a video, the motion of a face is the same as that of the real case and also gives the same life sign, such as photoplethysmography (PPG) signal,[8] lips movements, or eye blinking.[9]

The third type of face liveness detection is based on the 3-D facial information. It is easy to counterfeit the 2-D facial information; however, the depth information is quite hard to be counterfeited. Methods that obtain the depth information can be varied in a number of ways. Some attempts try to capture the face through different directions to manifest the 3-D information. Wang et al.[10] proposed that a real face can get different images when the shooting direction is varied. This method requires the user to rotate his head to get shootings from different directions, which is inconvenient, time-consuming, and can be fooled easily by the "video" type of attack that contains different directions shooting the user's face. Other works attempt to reveal the 3-D information

*Address all correspondence to Hong Fu, E-mail: hfu@chuhai.edu.hk

1017-9909/2019/$25.00

through evaluating the focusing degrees. Kim et al.[11] used a single camera with two different focusing lengths to get two images and Raja et al.[12] proposed to capture a stack of images with different focal lengths. Using these images to calculate, compare, and classify the real face and the facial image from spoofing attack. The performance of this kind of method is highly related to the distance between the camera and testing face. When the distance is going to be large, the performance degraded severely. Recently, light-field cameras (LFC) have been developed. Using light-field technique is very promising to solve face liveness detection problems and many other problems that need 3-D information, such as distance prediction, building modeling, object modeling, etc. We will introduce the attempts that are based on the light-field imaging after the brief introduction of the LFC in the next section.

LFC can provide depth information. We propose that this extra information can be well utilized by the convolutional neural networks (CNNs) to get better classification results for flat face spoofing attack. Because of this reason, we attempt to introduce CNNs into light-field-based liveness detection context. CNNs have recently got a great success in image or video classification, recognition, and retrieval since 2012.[13–17] At that year, Krizhevsky et al.[13] started an era in machine learning and CNNs, by activating a significant progress of the image recognition accuracy using CNNs. Almost every year after 2012, the score of ImageNet,[18] which has become the standard benchmark for large-scale object recognition in the past 7 years, has been improved a lot by CNNs-based method. Liveness detection also can be considered as an image classification problem. However, the difficulty of liveness detection is that all classes only have a slight difference, and for the light-field data, the data structure is not suitable for the CNNs model.

In this paper, we proposed two possible ways to convert the raw light-field data of the eye area into the type that is suitable for the conventional CNNs model. A conventional CNNs model was trained and used as the feature extraction and the classifier. In addition, we built a light-field face image database that includes the real face and three different types of the spoofing attack. All images in this database were taken by an LFC. The proposed system can get better performance on our database than the conventional image-based method and the existing light-field-based face liveness detection methods.

The rest of the paper is organized as follows: the related work is introduced in Sec. 2, which contains the light-field imaging principle, former light-field-based work, and the CNNs. Section 3 describes our proposed method. Section 4 describes the detail of the database that we have built, experimental setup, and the performance evaluation. Section 5 concludes this work.

2 Related Work

2.1 *Light-Field Camera*

The traditional imaging system [Fig. 1(a)] only records the 2-D information of the real-world scene by projecting it into a 2-D image sensor, which loses the object's depth information. To overcome this limitation, light-field imaging systems aim to collect both the total amount of light at each point on the photosensor and the amount of light traveling along each ray that intersects the sensor. To realize this, a lenslet-based light-field imaging system has been proposed by Adelson and Wang[19] and well implemented into a hand-held camera by Ng et al.[20] Figure 1(b) is a simple illustration of the microlens-based imaging system. An array of microlenses is placed at the image plane of the camera main lens. The image sensor is positioned in the focal plane of the microlens array. The lights come from the points on the plane of focus [the red color in Fig. 1(a)] and converge to a point on the image sensor in a traditional imaging system; however, in

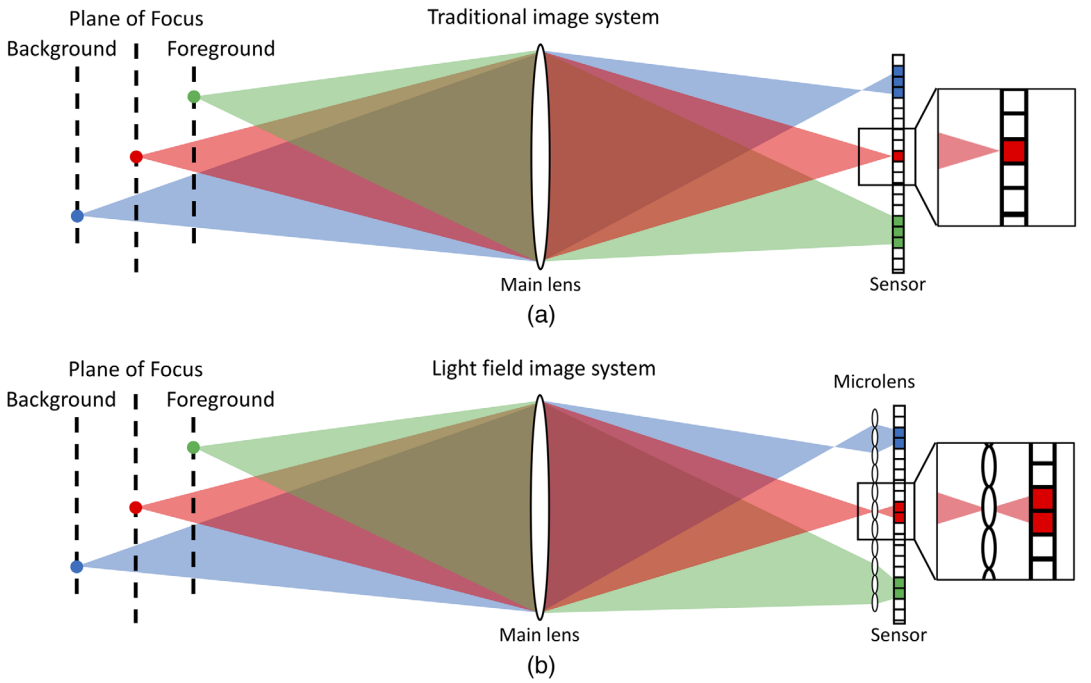

Fig. 1 The basic imaging principle comparison between (a) the traditional image system and (b) the light field image system.

a light-field imaging system, the microlenses at those points separate these light rays based on their directions [the red color in Fig. 1(b)]. The different part of the sensor will receive the light ray from different directions. Through this way, the focal plane can be changed after exposure and the depth information can be obtained from it.

2.2 *Light Field-Based Liveness Detection*

Sepas-Moghaddam et al.[21] gave a thorough review of the light-field-based face liveness detection recently, and following their work, we can classify the existing methods into two categories.

Approaches in the first category utilized the pixel intensity variance of different focus images. One important characteristic of the light-field imaging is that its focal length can be changed after one single shooting. Based on this reason, for the real face, the image may change a lot because the points on nose, eyes, or ears have a different distance to the camera. For example, if the focus area locates on the nose, the ears will become blurry. If the focus area moves to the left ear, the nose will become blurry. On the contrary, for the printed spoofing attack face or high-definition screen displayed spoofing attack, different parts of a face are on one plane. Wherever the focus area is located, the image should stay unchanged. Raghavendra et al.[22] have tried a lot of different methods based on this phenomenon to distinguish the *bona fide* face and spoofing attack face. They found that the Tenengrad variance can well reveal the difference between images with different focus areas. Ji et al.[23] proposed light field histogram of gradient (LFHOG) as the feature extracted from the light-field imaging data. LFHOG contains one more direction than the conventional HOG, which is applied to the depth direction. The multifocus image-based methods highly depend on the quality of the multifocus image. If the aperture parameter of the camera is small, the DoF should be very large. In this case, all the face area is in the DoF, which means wherever the focus area located, the image will not change a lot.

Approaches in the second category utilized the subaperture image of a light-field image data. The subaperture image is also the characteristic of the LFC. Because of the special structure of the LFC, the pixels on the photosensor are assigned to each microlens and form a small image, which is considered as the subaperture image.[21] Figure 2 is an example of the subaperture image. The number of pixels in one subaperture image relies on the number of the microlenses in the LFC. Kim et al.[24] concluded that the microlens image around the chin of the human face can distinguish the fake face and the real face, because, normally, the chin is close to the background, which leads to a large distance changing occurs in this area. This is the reason that this area has been chosen. However, this method is not robust to the complex background and the performance may decrease a lot when the subject's hair is too long to see their chin. Moreover, when the color of the background is similar to the face, the characteristic of the subaperture image in the chin area cannot significantly separate the real and the fake ones.

3 Proposed Liveness Detection method

The liveness detection is used as the pretest of the human face recognition-based security system, which aims to distinguish the fake face and the real face in front of the camera. Furthermore, the proposed system can also obtain the good ability on classifying the type of spoofing attack. Because of the imaging principle of the LFC, the raw data is a microlens image array after decoding, as shown in Fig. 2. The decoded data structure is a five-dimensional (5-D) data (u, v, x, y, c), where c represents the color channels, (x, y) is the pixel location, and (u, v) is the location in a microlens image that was highlighted in Fig. 2(c). We use the open source toolbox—light-field toolbox for MATLAB[25] to decode the raw data generated by the LFC. Based on this data structure and output, we design our proposed liveness detection method with the main pipeline shown in Fig. 3. First, the facial area will be detected on the plain image (u_0, v_0, x, y, c) generated from the 5-D data. (u_0, v_0) is the center location of a single subaperture image. This plain image has 3-D data structure that can be applied to the Viola et al.[26] face detection method directly. We used the OpenCV toolbox to implement this Viola–Jones face detection method. In addition, we also use the OpenCV toolbox to detect the eye area on each

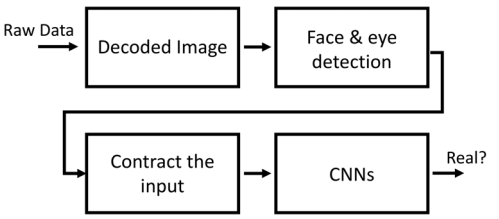

Fig. 3 The flow chart of the proposed liveness detection system.

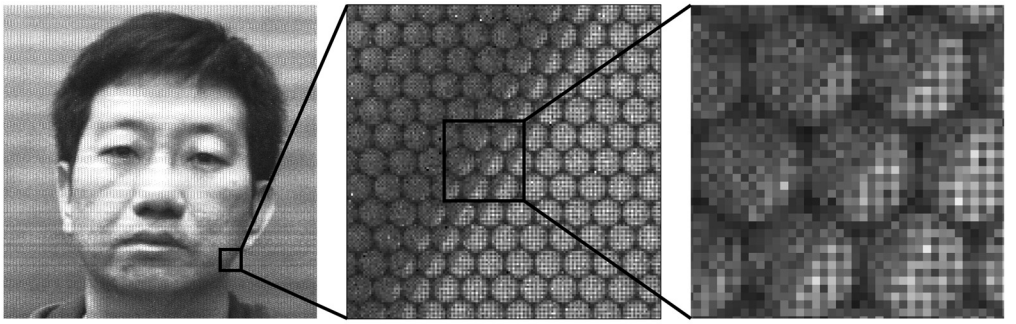

Fig. 2 An example of the light-field microlens images (before demosaicing) and the zoom in detail of the subaperture images.

detected face image to ensure that the detection result is reasonable. Both eye areas must be included in the face area. Second, we proposed two types of features—microlens image and ray difference image—to distinguish the real face and the spoofing attack face. Third, we propose to use CNNs to further extract the features and give the final detection result: real or spoofing attack in liveness detection or the type of attack in the spoofing attack classification.

3.1 *Microlens Images*

The size of the (u, v) domain that is generated by the Lytro Illum LFC is 15×15, which means every pixel in the plain image is a 15×15-pixel subaperture image [Fig. 4(c)]. Kim et al.[24] found that if the depth changed in an area, like the chin area of the face, the microlens image in that area should give an inhomogeneous pattern. For the real face, the depth always changes because the human face has irregular 3-D shape, but for the spoofing attack, normally it is either a 2-D image (printed attack or display attack) whose depth does not change too much or a regular 3-D image (wrapped photo attack) whose depth changes smoothly. The shape of the eyes and eye socket is complex and not easy to manipulate. A relatively large distance change usually appears between the upper eyelid and the eyeball. The pupil and sclera also have a slight distance difference and the color change is clear. All these characteristics only appear in the real case. The common ways of spoofing attack can only manipulate the color of eye area without any depth information. Based on this analysis, we designed an algorithm to build the microlens image that can maintain the pattern of the subaperture image and whose data structure (3-D) is suitable for the conventional CNNs.

Figure 4 shows the pipeline of the microlens image generation. The plain eye area image [Fig. 4(a)] is generated from the plain image (u_0, v_0, x, y, c), which selects the center pixel (u_0, v_0) of each subaperture image from the raw 5-D light-field data. Here, x and y are in the range of $[1, H]$ and $[1, W]$, where "H" and "W" are the height and width of the original image, respectively. We applied the $V - J$ algorithm in OpenCV toolbox to detect the eye area (u_0, v_0, x', y', c). x' and y' represent the location of this eye area and set the eye area a fixed size: 50×50-pixel. We found that the central area of each subaperture image, which contains the most key spatial information at that location in (x, y) domains, can already well maintain its pattern. Moreover, if we only consider the central area, the computational complexity can be reduced a lot. We chose to select the central 5×5-pixel area of each subaperture image to build the microlens image. At each location $((x_i', y_i'))$, the area $\{(u', v') : u' \in [u_0 - 2, u_0 + 2],\ v' \in [v_0 - 2, v_0 + 2]\}$ has been selected. The generated image (x'', y'', c) is a traditional plain image in which $x'' = x' \times 5 + u'$ and $y'' = y' \times 5 + v'$, which is suitable for the input of the conventional CNNs. The enlarged small patch in Fig. 4(b) can show the clear inhomogeneous pattern at the area between the eyelid and the eyeball and it also can be found between the pupil and the sclera.

3.2 *Ray Difference Images*

The second approach is named as the ray difference images. Every single pixel in a microlens image is considered as a direction of the light ray and also related to the focus plane.[19] The plain image that generated from the different position at a subaperture domain can be considered as the image is taken from the different focal lengths. For a 3-D object, such as the real face, the depth information of each point should be different, which leads to the different generated images. Following this principle, we designed the second type of feature. Figure 5 shows the basic pipeline of this approach. After the eye area has been detected in the same way as the microlens image, the location of the eye area, represented by the (x', y'), has been located. In the next step, the #1 to #4 in Fig. 5(b) and the #0 in Fig. 5(a) are generated by selecting different points on every subaperture images. The point #0 is the center point of the subaperture image and the other four points are the points that have

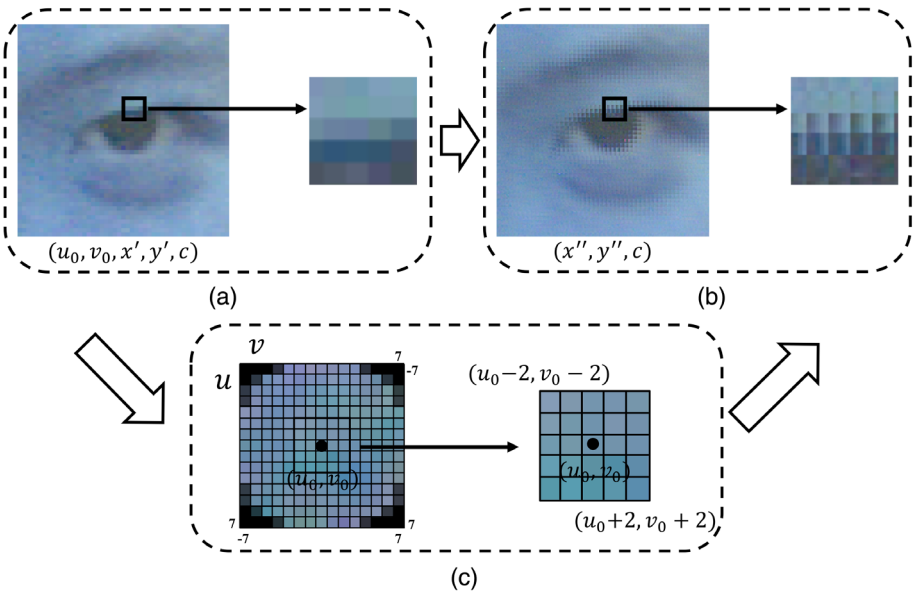

Fig. 4 Microlens image-based feature building. (a) An example of the detected eye area from a plain image and an enlarged small patch for illustration, (b) the generated microlens image and its enlarged small patch, and (c) the detail of the selected region in a subaperture image as an example.

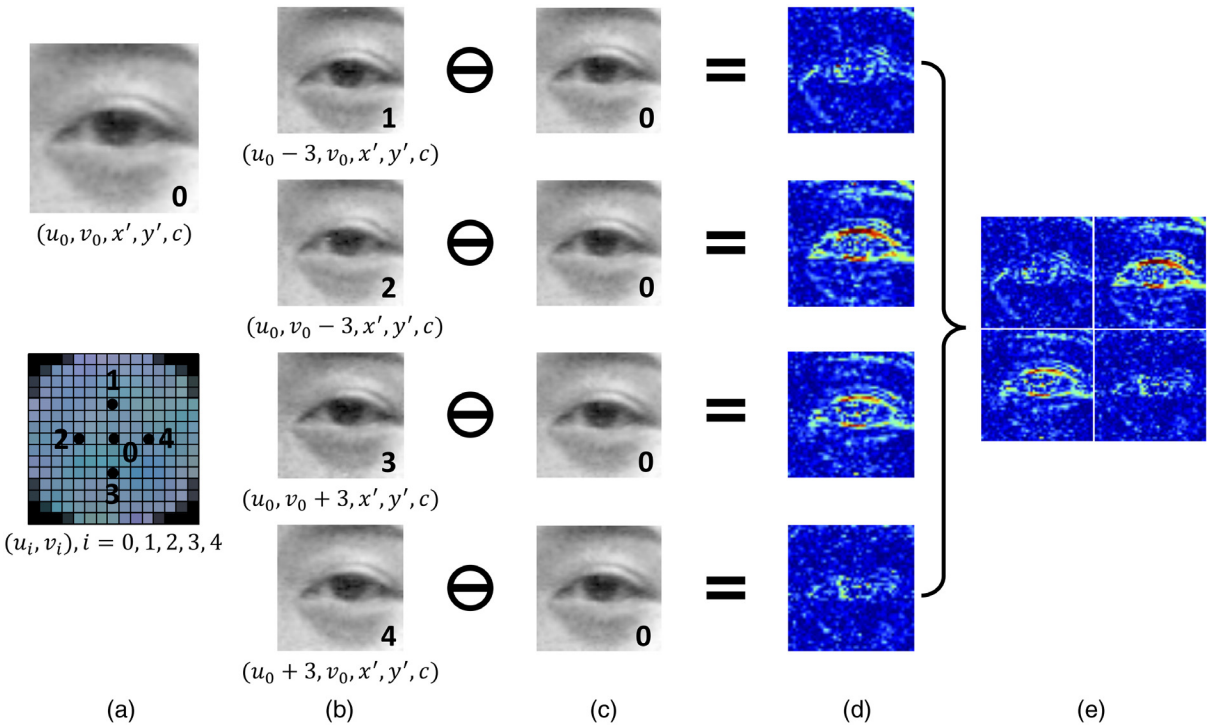

Fig. 5 Ray difference image based feature building: (a) the plain image generated from the central point of every subaperture images and the point selection on a subaperture image, (b) the plain images generated from the #1 to #4 points, (c) the (u_0, v_0, x', y', c) images, (d) the error maps between those generated images, and (e) concatenating all those error maps.

3-pixel distance from the central point in four different directions [Fig. 5(a)]. To reveal the difference of the generated image in Fig. 5(b), we designed to use those images to subtract the #0 image. Figure 5(d) shows the difference between those images and the central image. The area around the eyelid, which has depth change, has the largest difference—the lighter color in Fig. 5(d). This is an alternative way to emphasize the depth information using light-field imaging raw data. Finally, we concatenate those "difference" images to generate a single image as the input of the CNNs model.

3.3 *Structure of Convolutional Neural Networks*

The deep neural network we proposed contains four 2-D convolutional layers, two max-pooling layers, and two fully connected layers. Four convolutional layers contain 64, 128, 256, and 256 3×3 kernels, respectively. All the convolutional layers are activated by the ReLU function. The number of neuron of the fully connected layers is 256 and 128, respectively. One dropout with 50% keep probability is added between these two fully connected layers to avoid the overfitting during the training. The number of neurons at the final output layer is determined by the number of classes—two classes for liveness detection (real/fake) and four classes for spoofing attack classification (types of faces). The activation function used for the last layer is "softmax" [Eq. (1)], where $x^{(i)}$ and $\hat{y}^{(i)}$ are the input features and the predicted probability of the i'th training sample, respectively; "j" is the class index of this label; and θ_j. represents the trainable weights corresponding to class "j". The loss function is the cross-entropy function [Eq. (2)]. "$\hat{y}_n$" is the predicted probability of the n'th samples. y_n represents the ground truth class label of the current sample. We chose the "Adam"[27] with the default settings as the optimizer.

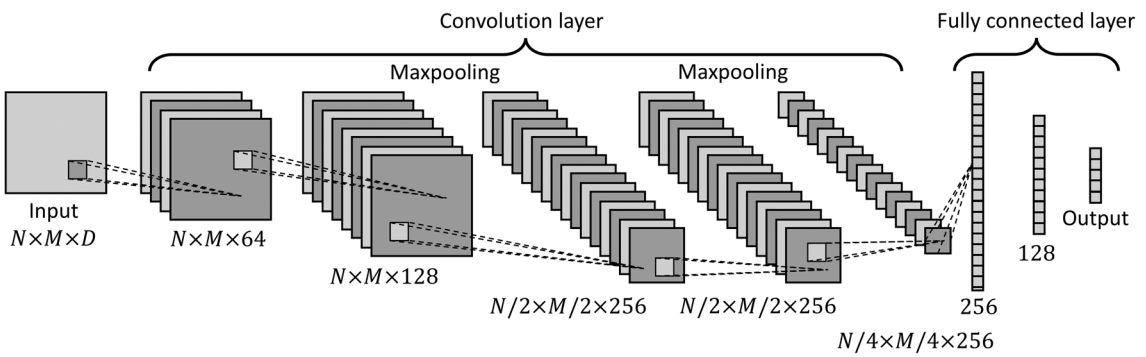

Fig. 6 Structure of the CNNs we used to further extract features and classify the different types of face.

Figure 6 is an illustration of the deep neural network structure:

$$p[\hat{y}^{(i)} = j|x^{(i)};\theta] = \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}, \quad (1)$$

$$L(\theta) = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)]. \quad (2)$$

4 Experimental Setup and the Results

To evaluate the performance of the proposed method, we build our own database with the human face images using Lytro ILLUM camera. As Lytro light-field cameras are no longer produced, Raytrix LFC can be an alternative for future research. This section introduces the detail about the database, the experimental settings, and the performance evaluation.

4.1 *Real and Spoofing Attack Face Database Construction*

This work mainly focuses on revealing the inherent characteristics of the LFC that can be used to separate the spoofing attack and real face in face recognition-based system. Our database contains 46 different people's faces with a homogeneous background. Each subject's photo is taken with different expressions to expand the size of the database. We design three different types of spoofing attacks—high-definition printed photo, warped printed photo, and a high-definition screen displayed photo. Photographs were taken at Chu Hai College of Higher Education over a period of 3 months. As all samples are volunteers and collected in college, 90% of the samples are young people whose age is between 20 and 29 years old. The gender ratio is around 1:1. The number of samples in real class and three different spoofing attack classes are 328, 199, 198, and 199, respectively.

The *bona fide* samples are captured with the homogeneous background. The subjects sit in a chair with a fixed height in front of the camera at 0.9 m. The lighting,

(a) (b) (c) (d)

Fig. 7 Illustrations of database collection: (a) real face, (b) printed photo, (c) warped photo, and (d) screen displayed photo.

(a) (b) (c) (d)

Fig. 8 Examples of artifact photo database which are collected by Lytro ILLUM camera and the face area is detected by Viola–Jones algorithm. (a) Original, (b) printed photo attack, (c) warped photo attack, and (d) screen displayed attack.

exposure, focal length, and other photography parameters are controlled. For spoofing attack dataset building (Fig. 7), we design three common types of the spoofing attack. Printed photo attack is very easy to realize and the 2-D facial information can be well reproduced by the printer machine. Warped printed photo attack contains depth information. The distance between different points of the warped face and the camera may be different. The HD screen displayed attack can represent two situations—photo or video is displayed. For the video based spoofing attack, many popular 2-D methods, such as motion based or live sign based, cannot perform well. Because the video can reproduce not only the motion of a human face in the real case but the life sign of a human being. The attack photo is printed by Fuji Xerox ApeosPort-V C7775 printer, which is colorful, high quality, and displayed by Dell P2214H 1920×1080 60-Hz screen. The printed and displayed photos are the original all-in-focus photos. Figure 8 shows the image examples of the database.[28]

4.2 *Experimental Settings and Results*

This part presents the experimental protocols that include the comparison methods, data structure, the data augmentation strategy, and the evaluation metrics. The Raghavendra et al.'s method[22] is considered as the baseline method. Two other well-developed 2-D methods (LBP[29] and Shearlet transform[30]) are also provided as the comparison methods. LBP is a well-known feature extraction, which is designed for representing the texture feature of the image. Reference 30 uses Shearlet transform as the feature extraction method and a fully connected neural network is used as the classifier. For the LBP-based methods, we design two types of inputs that are face and eye. The original design of LBP-based method uses the face image as input data, and as described in the previous sections, the eye area is worth to try as input data. The classifier of the LBP-based method is SVM. Because the eye area in our dataset is 50×50 pixel, which is too small to apply the Shearlet transform, we only implement this method on the face image. We also apply the conventional 2-D eye image extracted from our database as the input data on the same CNNs architecture to evaluate the effectiveness of the extra information introduced by the LFC. All experiments are implemented on the Linux Ubuntu 14.04 platform with python 2.7 language except the MATLAB toolbox to decode the light-field raw data. The CNNs is implemented on the tensorflow[31] platform.

About $250 \times 250 \times 3$ for microlens image and 100×100 for ray difference image are the dimensions of the CNNs' input and of the eye area is around 50×50 pixels, so we cropped each eye detection result to 50×50 pixels. Thus, for the microlens image, the dimension of CNNs' input is $250 \times 250 \times 3$, where 3 is the RGB three-color channels. The final ray difference image that was inputted to the CNN model obtains the $100 \times 100 \times 1$ dimension as it is stitched by four 50×50-pixels ray difference image. For LBP algorithm,[29] block-based multiscale LBP is used and both the face image and eye image are used as input. For the SBFD algorithm,[30] the Shearlet transform is applied on each 256×256-pixels grayscale face image. The input image is decomposed into four scales (exclude approximation component) and the direction number for each scale is 6. The pooling block size is 64 and the final SBFD length is 384. The classifier settings and training protocol of this method are the same as the original paper.[30] For all data driven methods, such as the proposed method, Litong's method,[30] and the SVM + LBP-based method, we randomly split the database into 80% samples for training and 20% samples for testing. In addition, the training set is further divided into 80% for training and 20% for validation. We found that the model is converged at around 50 epochs when the batch size is set to 32. Figure 9 shows the training process.

We designed a way to augment the data scale. We only applied the "shifting" on the decoded raw data because other common types of data augmentation on the conventional image or for recognition task are not suitable for the decoded raw light-field data and the liveness detection task. For example, changing the luminance value on the raw light-field data, such as adding salt and pepper or Gaussian noise and enlarging the average luminance value by multiplying a scale value, may totally change the characteristic of the decoded data. Moreover, the rotation translation matrix cannot be applied to both the raw data and the decoded data directly. To keep the data unchanged as much as possible, we only shifted the bounding box of the face or eye to generate the samples. The direction and the shift scale (%) are randomly selected in the range of [0, 360 deg] and [0, 10]. We ran the shifting operation 10 times to enlarge the scale of the samples 10 times. In addition, as the features we designed are based on the single eye, we consider two eyes from one subject as two samples with the same label. This should further enlarge the number of the samples twice.

We used the standard evaluation criteria to reveal and compare the performance of these methods. For liveness

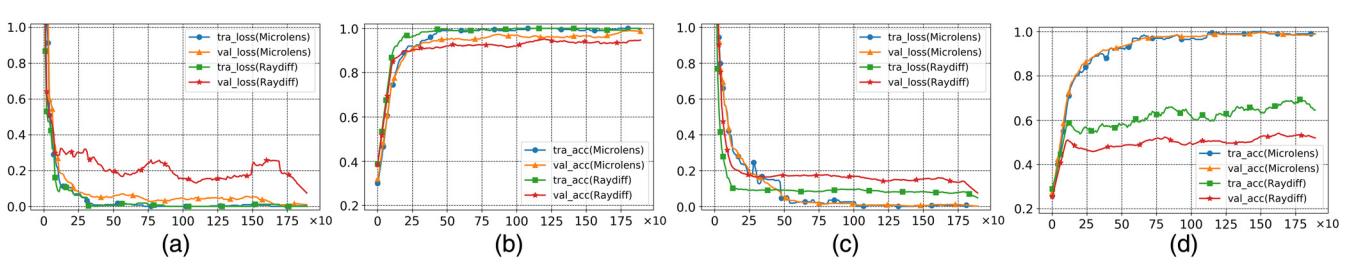

Fig. 9 Illustration of the training process in terms of the loss and the accuracy: (a) the changes of the loss value for the liveness detection task, (b) the changes of the accuracy value for the liveness detection task, (c) the changes of the loss value for the spoofing attack classification task, and (d) the changes of the accuracy value for the spoofing attack classification task.

detection, we followed the ISO/IEC WD 30107-3,[32] which is subject to three types of errors: *Bona Fide* presentation classification error rate (BPCER), the attack presentation classification error rate (APCER), and the average classification error (ACER). BPCER represents the proportion of the error of considering the true face as the spoofing attack; APCER represents the proportion of the error of considering the spoofing attack as a true face; and the ACER is the average value of the BPCER and the APCER. We chose the ROC curve to reveal the sensitivity of a liveness detection classifier, in which contains the false positive rate (APCER) as the x-axis and the true positive rate (one minus BPCER) as the y-axis. The curve that is closer to the upper left corner is better performance. For the spoofing attack classification, we use mean average precision (mAP)[33], which is the most popular parameter in multiclassification to evaluate the system property.

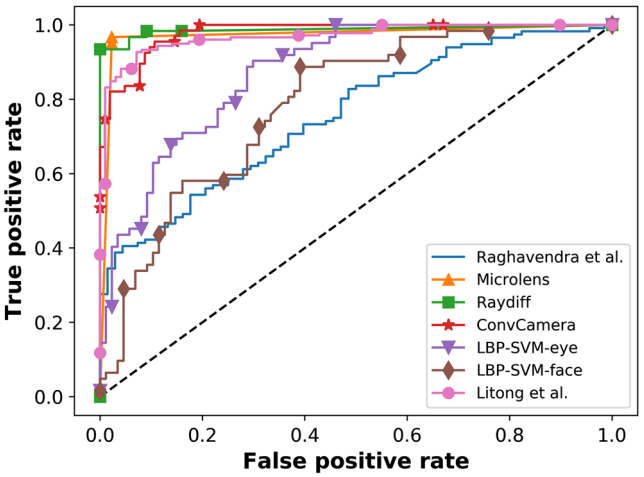

Fig. 10 ROC curve for different liveness detection methods.

4.2.1 *Evaluation of liveness detection*

The experimental results of liveness detection task are presented in this subsection. Table 1 shows the BPCER, APCER, and the ACER of each method. Figure 10 is the ROC curve of each method. For LBP-based methods, using the eye area image leads to better performance than using face image. This proves the previous analysis, which points out using the eye area is the better choice of feature extraction for liveness detection task than face area. Raghavendra's method cannot get good results from our database. This method focuses on the difference between the photos with different focal planes. We found that the LFC cannot set a very large aperture to get a narrow depth of field. As the depth of a human face is small compared with the depth of field, the difference between the images with the different focal point cannot be well revealed. "Litong" method performs lower ACER than other methods except for the proposed method. This result is the same in the original paper. The Shearlet transform can extract good features and the deep fully connected neural network takes advantage of these features to well classify the real and fake face. However, this method is not an end-to-end training; useful features may be ignored by the Shearlet transform. The proposed microlens-based methods outperform the previous methods. Especially, from the ACER value, both the proposed microlens-based and ray difference-based method can get much lower error rate than others. Moreover, we can find that the designed features introduced from the LFC can improve the performance of liveness detection by comparing with the case that using the conventional 2-D eye image as the input of the CNNs model ("ConvCamera" in Table 1) and the two proposed methods. The ROC curve in Fig. 10 also can show that the curve of two proposed methods is closer to the left upper corner than other methods, which indicates that the proposed methods can accept the real face with high accuracy when the accuracy for fake face rejection is also high.

4.2.2 *Evaluation of spoofing attack classification*

This subsection provides the results of spoofing attack classification. The spoofing attack classification aims to recognize the type of spoofing attack or the real face, which is useful for data collection for further security system design in the future. Table 2 shows that the proposed method can well recognize the different types of the faces by comparing the precision of each type of spoofing attack. The "LBP-eye" method performs the lowest mAP, which is not the same in the liveness detection task. The spoofing attack classification is a more challenging task that needs more information to

Table 1 Experimental results for liveness detection task (the bold values are the best performance).

Method	BPCER	APCER	ACER
Raghavendra	0.294	0.366	0.330
Litong	0.071	0.090	0.080
LBP-eye	0.321	0.127	0.224
LBP-face	0.378	0.213	0.296
ConvCamera	0.119	0.078	0.099
Microlens	**0.033**	**0.023**	**0.028**
Raydiff	0.066	0.023	0.045

Table 2 Average precision for spoofing attack classification (the bold values are the best performance).

Method	Real	Type1	Type2	Type3	mAP
Raghavendra	0.556	0.6	1	0.527	0.671
Litong	0.767	0.709	0.75	0.88	0.777
LBP-eye	0.604	0.357	0.85	0.7	0.628
LBP-face	0.718	0.563	0.641	0.938	0.715
ConvCamera	**0.919**	0.293	**0.905**	0.951	0.767
Microlens	0.908	**0.966**	0.903	**1**	**0.944**
Raydiff	0.868	0.722	0.808	0.675	0.768

classify the classes. The conventional 2-D image cannot well provide this information. Moreover, the high-quality printed photo can reproduce the face image very well, so the texture analysis-based methods, such as LBP base method, cannot perform well. The focus plane analysis method,[22] named "Raghavendra" in Table 2, can provide the accurate result for type 2 attack as this type attack has the depth changing that is the same with the real face; however, it cannot handle the other types of attacks very well. The "ConvCamera" has good performance for all types of attacks except the type 1 attack. The proposed microlens-based method still can get the highest mAP value. Although the precision of each class is not the highest, all of them are close to the highest value. The microlens image can well provide the discriminative information of the different classes and the CNNs model can utilize it well to get good performance.

5 Conclusion

This paper introduced a CNNs-based model as a classifier for LFC-based facial liveness detection. We designed two different features from the raw data of the light-field image: the microlens image and the ray difference image, as the raw light-field image data is not suitable for the existing CNNs model. Furthermore, we build a light-field image database for facial liveness detection. The experimental results show that our designed method—microlens method and CNNs can get lower ACER for the liveness detection (0.028 ACER) and higher precision for the spoofing attack classification (0.944 mAP) comparing with other current methods. Introducing CNNs in liveness detection, which is highly related to face recognition-based security access system, is valuable and promising to make liveness detection into the real product. The 3-D facial mask or mannequin attack has been raised, which can well manipulate the 3-D information of the whole face. The proposed method, which concentrates on the eye area, can be implemented to the iris liveness detection to further improve the robustness of the detection system in the future.

Acknowledgments

The work described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. UGC/IDS13/14). This work was also supported by the Key Research and Development Plan of Shandong Province (2018GGX101051 and 2018JHZ007).

References

1. W. Zhao et al., "Face recognition: a literature survey," *ACM Comput. Surv. (CSUR)* **35**(4), 399–458 (2003).
2. S. Thavalengal et al., "Iris liveness detection for next generation smartphones," *IEEE Trans. Consum. Electron.* **62**(2), 95–102 (2016).
3. D. Yi et al., "Face anti-spoofing: multi-spectral approach," in *Handbook of Biometric Anti-Spoofing*, pp. 83–102, Springer, London (2014).
4. R. S. Ghiass et al., "Infrared face recognition: a literature review," in *The 2013 Int. Joint Conf. on Neural Networks (IJCNN)*, IEEE, pp. 1–10 (2013).
5. O. Kähm and N. Damer, "2D face liveness detection: an overview," in *BIOSIG-Proc. of the Int. Conf. of the Biometrics Special Interest Group (BIOSIG)*, IEEE, pp. 1–12 (2012).
6. K. Kollreider, H. Fronthaler, and J. Bigun, "Evaluating liveness by face images and the structure tensor," in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, IEEE, pp. 75–80 (2005).
7. W. Bao et al., "A liveness detection method for face recognition based on optical flow field," in *Int. Conf. on Image Analysis and Signal Processing*, IEEE, pp. 233–236 (2009).
8. K. H. Suh and E. C. Lee, "Face liveness detection for face recognition based on cardiac features of skin color image," *Proc. SPIE* **10011**, 100110C (2016).
9. H. K. Jee, S. U. Jung, and J. H. Yoo, "Liveness detection for embedded face recognition system," *Int. J. Biol. Med. Sci.* **1**(4), 235–238 (2006).
10. T. Wang et al., "Face liveness detection using 3D structure recovered from a single camera," in *Int. Conf. on Biometrics (ICB)*, IEEE, pp. 1–6 (2013).
11. S. Kim et al., "Face liveness detection using variable focusing," in *Int. Conf. on Biometrics (ICB)*, IEEE, pp. 1–6 (2013).
12. K. B. Raja et al., "Robust face presentation attack detection on smartphones: an approach based on variable focus," in *IEEE Int. Joint Conf. on Biometrics (IJCB)*, IEEE, 651–658 (2017).
13. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012).
14. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *Lect. Notes Comput. Sci.* **8689**, 818–833 (2014).
15. P. Sermanet et al., "Overfeat: integrated recognition, localization and detection using convolutional networks," in *Int. Conf. on Learning Representations (ICLR2014)* (2014).
16. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3431–3440 (2015).
17. R. Girshick et al., "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(1), 142–158 (2016).
18. J. Deng et al., "ImageNet: a large-scale hierarchical image database," in *IEEE Conf. on Computer Vision and Pattern Recognition*, IEEE, 248–255 (2009).
19. E. H. Adelson and J. Y. A. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.* **14**(2), 99–106 (1992).
20. R. Ng et al., "Light field photography with a hand-held plenoptic camera," Computer Science Technical Report CSTR, vol. 2(11), pp. 1–11 (2005).
21. A. Sepas-Moghaddam, F. Pereira, and P. L. Correia, "Light field-based face presentation attack detection: reviewing, benchmarking and one step further," *IEEE Trans. Inf. Forensics Secur.* **13**(7), 1696–1709 (2018).
22. R. Raghavendra, K. B. Raja, and C. Busch, "Presentation attack detection for face recognition using light field camera," *IEEE Trans. Image Process.* **24**(3), 1060–1075 (2015).
23. Z. Ji, H. Zhu, and Q. Wang, "LFHOG: a discriminative descriptor for live face detection from light field image," in *IEEE Int. Conf. on Image Processing (ICIP)*, IEEE, pp. 1474–1478 (2016).
24. S. Kim, Y. Ban, and S. Lee, "Face liveness detection using a light field camera," *Sensors* **14**(12), 22471–22499 (2014).
25. D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1027–1034 (2013).
26. P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vision* **63**(2), 153–161 (2005).
27. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Int. Conf. on Learning Representations (ICLR)* (2015).
28. H. Fu, "Light field-based face spoofing attack dataset (LF-SAD)," https://sstrc.chuhai.edu.hk/index.php/research-project/ (23 October 2018).
29. T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002).
30. L. M. Po et al., "Face liveness detection using shearlet-based feature descriptors," *J. Electron. Imaging* **25**(4), 043014 (2016).
31. M. Abadi et al., "Tensorflow: a system for large-scale machine learning," in *Proc. of the 12th USENIX Conf. on Operating Systems Design and Implementation (OSDI'16)*, Vol. **16**, pp. 265–283 (2016).
32. ISO/IEC JTC1 SC37 Biometrics, "Information technology-presentation attack detection—Part 3: testing, reporting and classification of attacks, document ISO/IEC 30107-3:2017," International Organization for Standardization (2017).
33. E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Inf. Process. Manage.* **36**(5), 697–716 (2000).

Mengyang Liu is working toward the PhD at the Department of Electronic Engineering, City University of Hong Kong. He received his BE degree in optoelectronic engineering and his MSc degree with dissertation in electronic and information engineering from City University of Hong Kong, China, in 2014 and 2015, respectively. His research interests include image and video processing, video indexing, computer vision, and machine learning.

Hong Fu received her bachelor's and master's degrees from Xi'an Jiaotong University in 2000 and 2003, respectively and her PhD from The Hong Kong Polytechnic University in 2007. She is now an associate professor in the Department of Computer Science, Chu Hai College of Higher Education, Hong Kong. Her research interests include pattern recognition, artificial intelligence, and their applications.

Ying Wei received her BS and MS degrees from Xi'an Jiao tong University, China, in 2000 and 2003, respectively. She received her PhD from National University of Singapore in 2008. Currently, she is a professor at the School of Control Science and Engineering in Shandong University. Her research interests include digital filter design, high-speed digital systems, and biomedical signal processing. She serves as an associate editor of *IEEE Transactions on Biomedical Circuits and Systems*.

Yasar Abbas Ur Rehman received his BSc degree in electrical engineering from City University of Science and Information Technology, Peshawar, Pakistan, in 2012, and his MSc degree in electrical engineering from the National University of Computer and Emerging Sciences, Pakistan, in 2015. Currently, he is pursuing his PhD in the Department of Electronic Engineering, City University of Hong Kong. His research interests include deep learning, computer vision, machine learning, face anti-spoofing techniques and wireless multimedia sensor networks (WMSNs).

Lai-Man Po received his BS and PhD degrees in electronic engineering from City University of Hong Kong, Hong Kong, in 1988 and 1991, respectively. He has been with the Department of Electronic Engineering, City University of Hong Kong, since 1991, where he is currently an associate professor. He has authored over 150 technical papers. His research interests include video processing and deep learning with an emphasis on face recognition, and image quality assessment algorithms.

Wai-Lun Lo received his BEng and PhD degrees in electrical engineering from the Hong Kong Polytechnic University, Kowloon, in 1991 and 1996, respectively. Currently, he is the head of the Department of Computer Science, Chu Hai College of Higher Education and the advisor for the Centre for Smart Energy Conversion and Utilization Research, College of Science and Engineering, City University of Hong Kong. His research interests include computational intelligence and applications of machine learning.