Pattern Recognition 83 (2018) 340–352

Contents lists available at ScienceDirect

Pattern Recognition

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

Visual speaker authentication with random prompt texts by a dual-task CNN framework

Feng Cheng[a], Shi-Lin Wang[a,*], Alan Wee-Chung Liew[b]

[a] School of Cyber Security, Shanghai Jiaotong University, Shanghai, China
[b] School of Information and Communication Technology, Griffith University, Gold Coast Campus, Queensland QLD4222, Australia

ARTICLE INFO

Article history:
Received 4 November 2017
Revised 18 May 2018
Accepted 6 June 2018
Available online 20 June 2018

Keywords:
Visual speaker authentication
Deep convolutional neural network
Multi-tasks learning
Liveness detection

ABSTRACT

Good authentication performance and liveness detection are two key requirements in many authentication systems. To avoid replay attacks, a novel visual speaker authentication scheme with random prompt texts is proposed. Compared with the fixed password scenario, visual speaker authentication with random prompt texts is much more challenging because it is impossible to ask the client to pronounce every possible prompt text to be used as training samples. In order to solve this problem, a new deep convolutional neural network is proposed in this paper and it has three functional parts, namely, the lip feature network, the identity network, and the content network. In the lip feature network, a series of 3D residual units have been adopted, which can depict the static and dynamic characteristics of the lip biometrics comprehensively. By considering the distinguishing features of the identity and content authentication tasks, the identity network and the content network are designed accordingly. An end-to-end, multi-task learning scheme is proposed which can optimize the weights of all the above three networks simultaneously. Experiments have been carried out to evaluate the performance of the proposed network under both the fixed-password and the random prompt texts scenario. From the experimental results, it is shown that the proposed approach can achieve superior performance in the fixed-password scenario compared with several state-of-the-art approaches. Furthermore, it also achieves satisfactory authentication results in the random prompt texts scenario and thus it provides a reliable solution for user authentication where liveness is guaranteed.

1. Introduction

In recent years, biometric feature based person authentication systems have attracted more and more research interests. Compared with the traditional approaches using Personal Identity Number (PIN) or passwords, biometric feature based authentication approaches can achieve higher security and better convenience. Human face and fingerprint are widely used biometric features for user identity authentication in many access control systems. However, these systems may be compromised by artificial fingerprints [1], pre-recorded videos or still face images [2]. Hence, liveness detection has become a challenging task in most biometric feature based authentication approaches.

Recent research has shown that human lip and its movement during utterance contain a lot of identity related information and thus it can be regarded as a new kind of biometric feature for person authentication [10–21,35]. Note that the lip feature is a twin biometrics which contains both physiological (different lip shape and appearance) and behavioral information (different behavioral pattern caused by speaker's talking habit). Compared with the traditional biometric features, lip biometrics has the following advantages: i) the dynamic characteristics of the lip biometrics guarantees a high level of security as well as robustness. On one hand, imposters would have great difficulty in imitating a user's talking habits. On the other hand, by assigning a different prompt text during each verification, it is also very effective to resist replay attacks and guarantee liveness; ii) the feature capturing devices, i.e. various kinds of video cameras, are easily accessible and the capturing process is less intrusive compared with those of the iris or fingerprint features; iii) the lip feature can be easily integrated with face, voice, etc. to construct a multimodal biometric system to provide a very high level of security [3]. In view of the above, lip biometrics has a good potential for user acceptance and industrial practicability and it has attracted much research interests in the past decade.

Generally speaking, the main challenge of using lip biometrics in user authentication is how to extract effective features from the lip image sequences. The desired features need to be of high

* Corresponding author.
E-mail addresses: klaus.cheng@qq.com (F. Cheng), wsl@sjtu.edu.cn (S.-L. Wang), a.liew@griffith.edu.au (A.W.-C. Liew).

https://doi.org/10.1016/j.patcog.2018.06.005
0031-3203/

discriminative power in differentiating different speakers and also robust to variations caused by a speaker's pose and distance towards the camera, etc. Furthermore, how to protect the authentication system against possible spoofing attacks such as replaying prerecorded videos is another critical issue. In this paper, a new speaker authentication system based on the lip feature is proposed, which takes both the authentication accuracy and robustness against replay attacks into consideration. In the authentication stage, the user is asked to pronounce a prompt text randomly generated by the system and our authentication system will check whether he/she is the right person and whether he/she is pronouncing the correct sentence simultaneously. The above mechanism can effectively protect against replay attacks but it brings additional requirements for the authentication system, which involves verifying both the speaker's identity and the speech content. In order to solve this problem, we have designed a deep convolutional neural network (DCNN) based algorithm, which considers both the identity and content information. DCNN is an important network structure in deep learning, which can automatically extract inherent and discriminative features from the training samples and has shown great success in image/video classification [4,5], human pose recovery [6,7] and human pose recognition [8,9]. The major contributions of the proposed algorithm can be summarized as follows. First, we propose a novel lip feature representation based on the spatiotemporal DCNN, which is referred to as the lip feature network. By introducing the data augmentation procedure, the extracted DCNN features are robust against variations caused by scaling, rotation and translation and thus can handle variation in speaker's pose and distance towards the camera. To our best knowledge, it is the first DCNN-based feature representation for visual speaker authentication. Second, considering the inherent relationship between the lip-based speaker recognition and speech recognition problems, a dual-task learning scheme is proposed. Two DCNN-based networks for speaker authentication and speech authentication are designed, which share the same input from the lip feature network. Third, the proposed DCNN network is an end-to-end model which integrates the functions of feature extraction from the lip image sequences, identity authentication, and content authentication. Hence, no pre-processing and post-processing procedure is required.

The rest of this paper is organized as follows. Section 2 briefly overviews the relevant works. Section 3 presents the proposed DCNN framework for visual speaker authentication, where each module is elaborated in detail. Section 4 presents the experimental results of the proposed scheme in comparison with three state-of-the-art approaches. Finally, Section 5 draws the conclusion.

2. Relevant works

Lip biometrics as a means of identity authentication, which is also referred to as visual speaker authentication, was first introduced in [22]. During the past decades, many researchers have proposed various approaches in this area [10–21]. Generally speaking, the existing methods can be roughly divided into two categories, i.e. model-based approach [10–13,15,17,19] and region-based approach [14,16,18,20,21]. Table 1 provides an overview of various visual speaker authentication approaches solely using the lip feature.

For the model-based visual speaker authentication approach, a lip model has to be designed and automatically extracted from the facial images containing the lip region. In [10], Wark et al. adopted the active shape model (ASM) to build the lip model, and the model parameters depicting the lip shape and intensity variation around the lip contour are adopted as the lip features. After feature extraction by discriminant analysis, the Gaussian Mixture Model (GMM) was adopted for classification. Their approach can successfully differentiate all the speakers in the TULIPS database [23] with twelve speakers. Jourlin et al. [11] also employed the ASM model parameters as the lip feature and the hidden Markov model (HMM) for classification. A Half Total Error Rate (HTER) [25] of 15.4% was achieved in the M2VTS database [24] with 37 speakers.

In [12], Broun et al. combined the teeth and tongue information with the lip shape features and they achieved a HTER of 6.3% on the XM2VTS database [25] (261 speakers out of 295 were selected). Centigul et al. showed in [13] that the motion vector of the lip region contained useful information to differentiate different speakers. By discrimination analysis, their approach can achieve an Equal Error Rate (EER) of 5.2% in the MVGL-AVD database with 50 people. In [15], Sanchez and Kittler tried to fuse lip feature with the face and voice biometrics. Dynamic Time Warping (DTW) was adopted as the classifier and a HTER of 13.35% was achieved on the XM2VTS database with 295 clients using only the lip feature. Liu et al. proposed a series of delicate shape features for visual speaker verification [17] and a HTER of 7.11% was obtained on the Cohn–Kanade database with 96 speakers. In our previous work [19], a comprehensive study was performed on the physiology and behavioral part of the lip biometrics and the optimal static and dynamic lip feature sets were then proposed. An EER of 1.92 was achieved on a database containing 40 speakers. The authentication performance of the model-based approaches greatly depends on the accuracy and robustness of the lip model. Moreover, additional computations are required to fit the lip model to the lip image.

On the contrary, region-based visual speaker authentication approach usually works directly on a rough lip region instead of an accuracy lip model. Various kinds of feature representations have been proposed to describe the static and dynamic information of the lip region. In [14], Faraj and Bigun adopted velocity estimation to describe the dynamics of the lip region. The centroids of the velocity vectors were used as the lip feature. They achieved an EER of 22% on the XM2VTS database. Samad et al. [16] have demonstrated the feasibility of visual speaker authentication using the static lip information only. Using their approach, ten speakers in the AMP-CMU database were authenticated successfully. Liu et al. [20] performed PCA and 2D-DCT on the lip region and extracted corresponding features. Then a multi-boosted HMM scheme was proposed to extract the most discriminative segment in the utterance of a specific user. An EER of 4.6% was achieved in a database with 46 speakers. In 2012, Chan et al. [18] proposed a local texture descriptor, i.e. the Local Ordinal Contrast Pattern (LOCP), to describe the lip region. The Three Orthogonal Planes (TOP) was also employed to combine the spatial and temporal information. They performed experiments on the XM2VTS database and achieved a very low HTER of 0.36%. Recently, our group has proposed a sparse coding based lip feature representation to describe the lip region and its dynamics in a spatiotemporal manner [21]. Experiments on a database with 40 speakers have demonstrated the effectiveness of the proposed feature (with an EER of 0% and HTER of 0.46% using limited training samples). As a result, region-based approach can achieve reliable authentication results without knowing the exact lip model; however, variations caused by different talking poses and positions will degrade the performance to some extent [21].

Most of the existing techniques assumed that the password for each user is fixed during training and authentication. Such assumption ensures a double security, i.e. the password and the talking behavior; however, a replay attack using a prerecorded video can deceive the authentication system. In order to perform identity authentication with liveness detection, an authentication scheme with random prompt texts (or passwords) was proposed in [26] and the flowchart of the scheme is given in Fig. 1.

The random prompt texts scheme can effectively protect the system from replay attacks. However, since different prompt texts

Table 1
A brief overview of existing lip-based authentication systems.

Method	Category	Database	#Clients	Performance
Wark (1997) [10]	Model-based	TULIPS	12	EER = 0%
Jourlin (1997) [11]	Model-based	M2VTS	37	HTER = 15.4%
Broun (2002) [12]	Model-based	XM2VTS	261	HTER = 6.3%
Centingul (2006) [13]	Model-based	MVGL-AVD	50	EER = 5.2%
Faraj (2006) [14]	Region-based	XM2VTS	295	EER = 22%
Sanchez(2006) [15]	Model-based	XM2VTS	295	HTER = 13.35%
Samad (2007) [16]	Region-based	AMP CMU	10	HTER = 0%
Liu (2012) [17]	Model-based	Cohn–Kanade	96	HTER = 7.11%
Chan (2012) [18]	Region-based	XM2VTS	295	HTER = 0.36%
Wang (2012) [19]	Model-based	Their Own	40	EER = 1.92%
Liu (2014) [20]	Region-based and Model-based	Their Own	46	EER = 4.6%
Lai (2016) [21]	Region-based	Their Own	40	EER = 0.00%, HTER = 0.46%

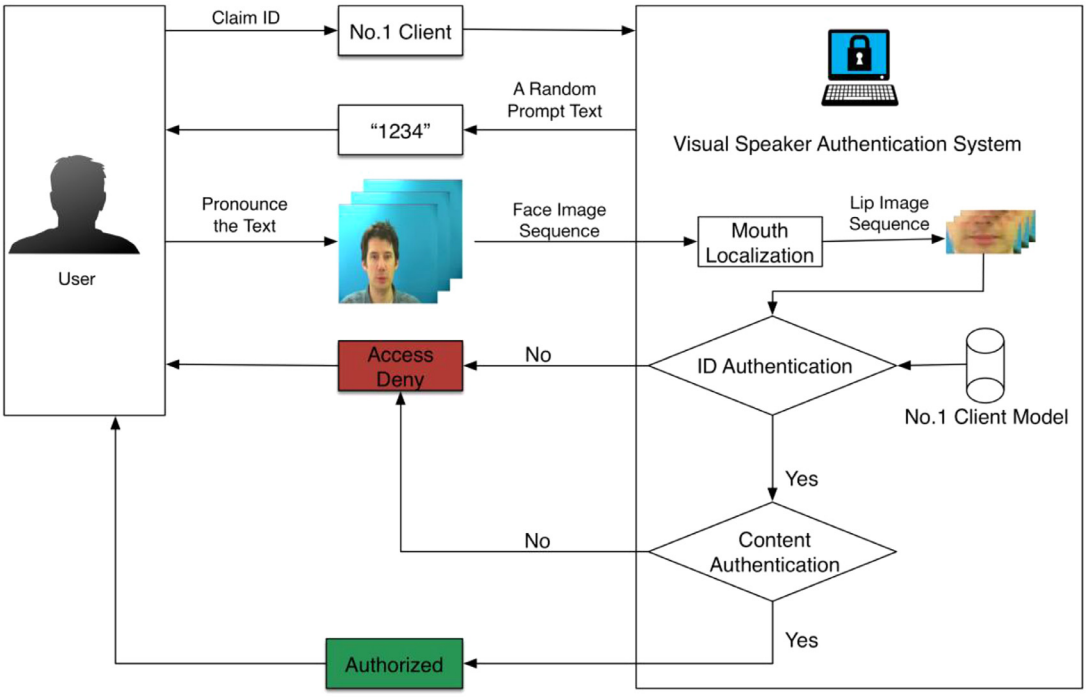

Fig. 1. The flowchart of the system.

rather than a fixed password are used in the authentication stage, most existing lip-based authentication systems have great difficulty implementing such scheme directly. The major problem lies in the training stage. For the traditional approaches with a fixed password, the user was asked to repeat his/her password several times and then the authentication system can build a specific model to describe his/her talking style when pronouncing the password. When the number of prompt texts or password increases, the number of required training samples increases accordingly and the user will become impatient if he/she is asked to pronounce so many utterances during training. In order to solve this problem, we make an assumption that the random prompt texts contain limited vocabulary. For example, the prompt text set contains all the four-digit phrases from "0000″ to "9999″, where the vocabulary is from 0 to 9. In traditional authentication systems using fixed password, each four-digit phrase needs to be repeated for several (e.g. three) times and the overall training set contains 3*10,000 = 30,000 phrases. In our system, we only require that each word (i.e. digit from 0–9) in the vocabulary be repeated several times during training. For example, if we have the training phrases consist of "1234", "5678" and "9090", which cover each of the 10 digits at least once, and we require that each digit to be repeated 3 times, the overall training set would contain 3*3 = 9 phrases. This greatly reduces the required number of training samples and improves the feasibility of the authentication system. Since most prompt texts have not appeared in the training set, it is very challenging for the authentication system to correctly authenticate them. In [26], the authors trained models for each word in the prompt texts (e.g. all the ten digits) and get acceptable authentication results for synthesized sequences. However, such kind of approach cannot handle well the variation caused by coarticulation of sounds and is not robust against replay attacks. As far as we know, there is no approach in the literature that can adequately handle the problem of speaker authentication in the continuous speech, random prompt texts scenario.

3. The proposed method

In our approach, a new deep convolutional neural network (DCNN) has been proposed for visual speaker authentication with random prompt texts. The major advantages of DCNN in this application lie in the followings. First, DCNN is a powerful feature representation tool, which can comprehensively describe the static and dynamic characteristics of the lip feature. Meanwhile, it can also extract useful features to differentiate various speakers and speech

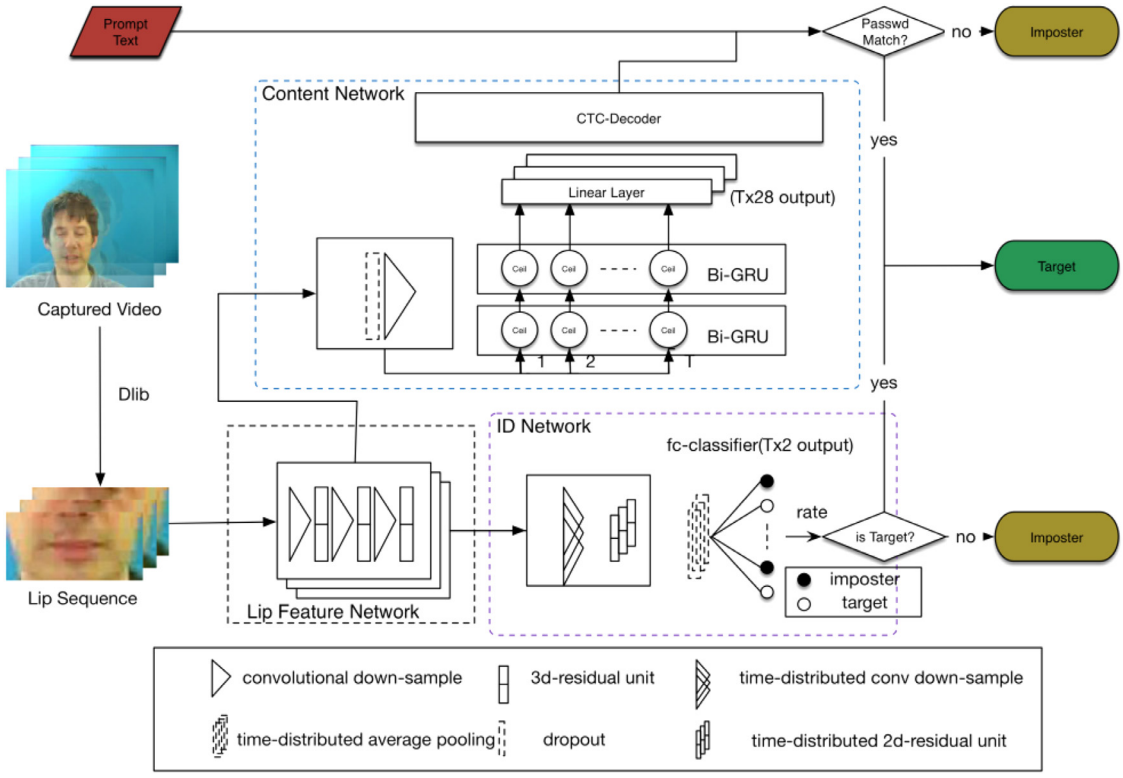

Fig. 2. The overall architecture of our system.

contents. Second, DCNN provides a flexible architecture, which can integrate the functionality of identity authentication and content recognition in one network. Third, with DCNN, end-to-end training can be realized and the inherent relationship between the original lip image sequence and the speaker's identity & the speech content can be learned simultaneously and directly from the training samples. By considering the characteristics of the lip feature, the network structure of the proposed method is designed accordingly and the details are elaborated in the following subsections.

3.1. Overview of the authentication system

Fig. 2 shows the overall architecture of the proposed authentication system. After generating a random prompt text, the system captures the frontal face image sequences using a video camera when the user pronounces the prompt text. Then the Dlib face detector, which is a face landmark detector with an ensemble of regression trees [27], is used to localize and extract the lip region from each captured face image. The extracted lip image sequences are then used as the input to the proposed DCNN (which is referred to as the Lip-Authentication-Network, LAN in short, hereafter). Generally speaking, our LAN can be roughly divided into three parts, i.e. the Lip-Feature-Network (LF-Net in short), the Identity-Network (ID-Net in short) and the Content-Network (C-Net in short). The LF-Net aims to extract the lip features (output) from the lip image sequences (input) and these features are expected to be representative and comprehensive in describing the static and dynamic information of the lip region during utterances. Following the LF-Net, the ID-Net and the C-Net are designed to predict the speaker's identity and the speech content from the lip features, respectively. The final authentication result is determined by the outputs of both the ID-Net and the C-Net. In the following sub-sections, the details of the LF-Net, the ID-Net and the C-Net will be elaborated.

3.2. The LF-Net

For both visual speaker authentication and visual speech recognition systems, how to extract representative and discriminative lip features has always been a very critical issue. Considering that the lip dynamics are of great importance in both identity authentication and speech content recognition, three-dimensional convolutional kernels are adopted to capture the spatiotemporal characteristics of the lip dynamics. Moreover, inspired by the ResNet [6] and TDNN [28], a series of 3D-convolutional layers are stacked together to provide a hierarchical representation, which can comprehensively describe the local and global details of the lip region and its movements.

The network structure of the proposed LF-Net is shown in Fig. 3. There are three prominent features of the proposed structure, which makes it suitable for lip feature representation. First, in each 3D convolutional downsampling layer, the spatial resolution has been reduced into half in both the horizontal and vertical directions, which enables the LF-Net to provide a hierarchical representation of the lip image. Second, the length of the convolutional kernel in the temporal domain is set to 3 and the stride is set to 1 in all the convolutional layers. Hence, after each 3D convolutional layer, the output features at time t are determined by the corresponding input values at time t-1, t & $t+1$.

From Fig. 4, it is observed that the reception field of the extracted features at time t in the i-th layer is from $-i$ to $t+i$. Considering that the proposed LF-Net is composed of three 3D residual units with an overall of nine 3D convolutional layers, the final feature at time t can capture the static and dynamic information of the current and the neighboring $2\times 9=18$ frames. It is interesting to note that since the frame rate of the video camera is usually set to 25–30 frames/second, nineteen frames in the lip image sequence can usually represent the lip information during the utterance of a single word. Hence, the final feature is able to provide the lip

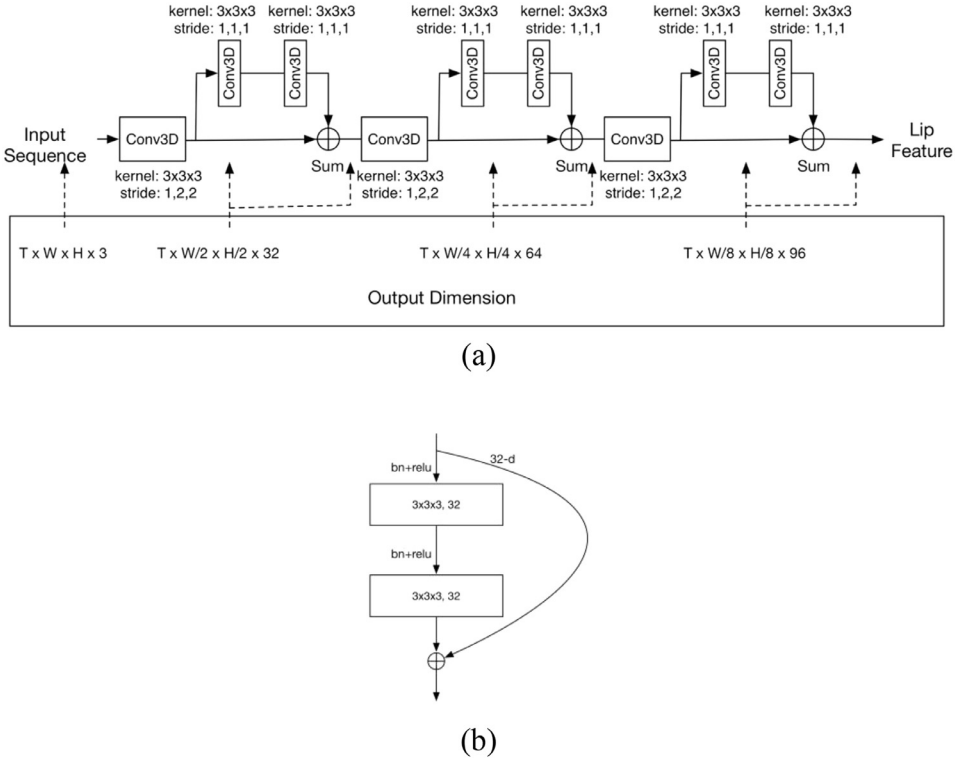

Fig. 3. (a) The overall network structure of the LF-Net; (b) A 3D residual unit with the input feature dimension of 32.

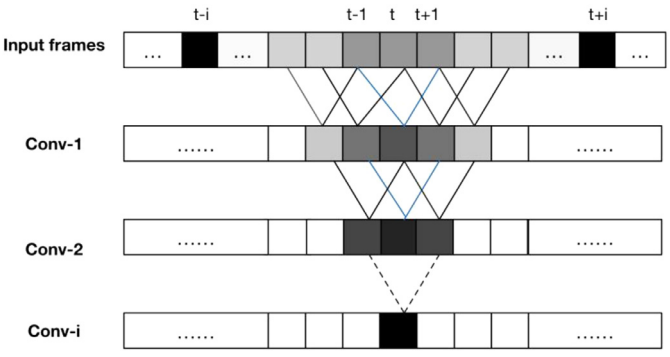

Fig. 4. The reception field of 3D convolutional layers in the time domain.

dynamics at the word level, which can well serve the subsequent ID-Net and C-Net. Third, by introducing the shortcut, the network can achieve high feature representative power and high robustness against overfitting.

Fig. 5 shows the saliency map of the outputs of the LF-Net from a lip image sequence. From the figure, it can be seen that most of the saliency pixels are gathered around the lip region (esp. the lower lip region), which demonstrates that the proposed LF-Net can automatically extract identity and content related information from the lip image.

3.3. The ID-Net

The ID-Net aims to design and construct an inherent relationship between the fundamental lip features and the speaker's identity. As discussed in our previous work [19], the lip feature is a twin biometrics and both the physiological (static) and behavioral (dynamic) information are useful in indicating the speaker's identity. Moreover, as indicated in [17,18,33], the user's unique talking style or mannerism can be reflected in short segments when he/she is pronouncing a word, a phoneme or even a series of sounds. Hence, in the proposed ID-Net, only the short-period lip features rather than those representing the entire lip sequences are taken into consideration.

The network structure of the ID-Net is shown in Fig. 6. Recall that the output features of the Lip-Net at each time spec contain static and dynamic information of the neighbouring eighteen frames, which can adequately describe the variations of the lip region in such a short period. The ID-Net is then designed to learn the inherent relationship between the fundamental lip features and the speaker's identity. Note that in the ID-Net, the features at different time spec are treated independently because our objective is to identify local user-specific talking styles regardless of the pronunciation order, which can also make the ID-Net insensitive to different speech content. Hence, the time-distributed (TD) convolutional layers have been employed and the input features at different time specs share the same 2D convolutional kernel in the spatial domain. In order to increase the feature representation power, one 2D residual unit is also incorporated between the time-distributed convolutional layer and the time-distributed average pooling layer. Finally, two fully connected layers with a softmax output have been adopted for classification.

3.4. The C-Net

The C-Net aims to authenticate the speech content, i.e. whether the speaker has pronounced the correct prompt text. In contrast to the identity-related features, which are usually contained in short segments, the content-related features are usually "global", i.e. the context information is of vital importance in speech content recognition. Especially in our application, the possible prompt texts are of limited vocabulary and have specific rules in texts generation, e.g. the prompt text can be composed of four digits (from "0000" to "9999"). In such scenario, the context information can provide more useful information in identifying which prompt text has been

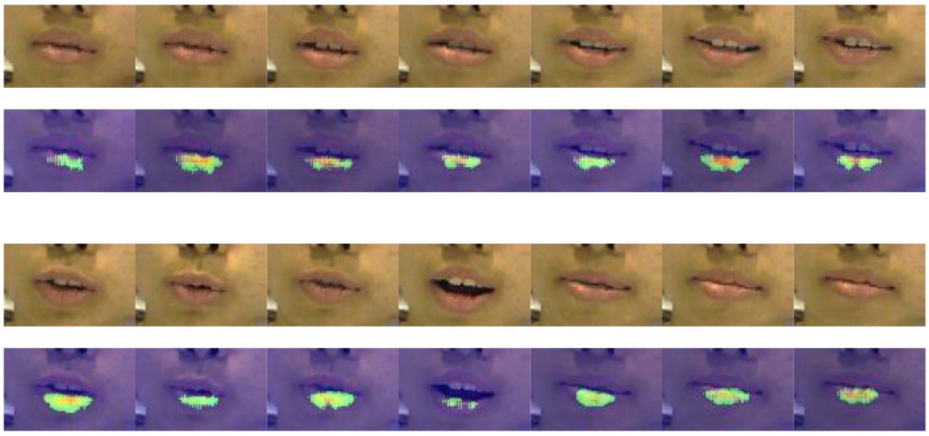

Fig. 5. Saliency map of lip image sequences.

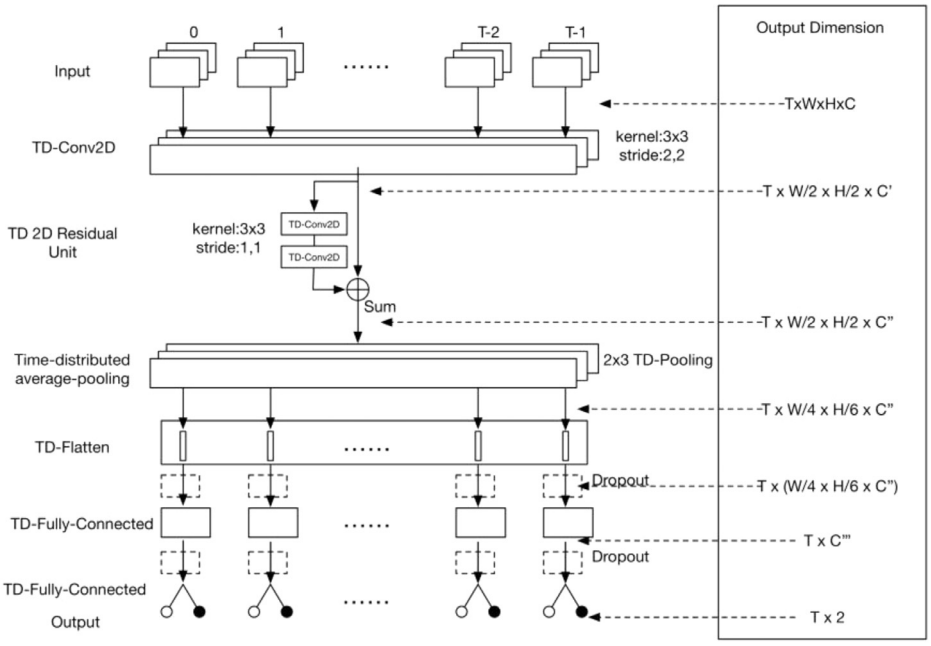

Fig. 6. The network structure of the ID-Net.

pronounced. In view of the above and inspired by the Lip-Net [29], the recurrent neural network layers have been adopted in the proposed C-Net.

Fig. 7 shows the network structure of the proposed C-Net. A 3D convolutional layer has been applied to extract linguistic related information from the fundamental lip feature sequences. Then, the output features have been time distributed flattened and input to the bi-directional recurrent neural network layers. Similar to the Lip-Net [29], the Gate Recurrent Unit (GRU) [30] has been adopted as the basic unit due to its better performance and faster convergence compared with that of the Long Short Time Memory (LSTM) [31]. A time distributed fully connected layer has been employed to extract content-related features from the output of the bi-directional GRU layers and the Connectionist Temporal Classification (CTC) [32] has been adopted for content recognition, which has shown superior performance in speech recognition and can well handle the variations caused by time delay and speaking speed, etc. Moreover, CTC can achieve end-to-end learning without any additional preprocessing steps such as word segmentation.

3.5. Dual-Task learning

Multi-Task Learning (MTL) [33] is a widely used machine learning technique where a series of tasks with certain commonalities and differences are learned together. In our application, there are overall two tasks, i.e. to authenticate the speaker's identity and to recognize the speech contents from the lip image sequences. The LF-Net is designed to describe their shared commonalities, i.e. to extract useful features which can describe the static and dynamic information of the lip region comprehensively. The ID-Net and C–Net are designed to fulfill their different tasks respectively. A dual-task learning approach is proposed, which can optimize all the above three networks simultaneously. It runs as follows and the sketch is given in Fig. 8.

The overall loss is formulated in Eqn. (1)–(3), i.e.,

$$L = \nu L_{ID-Net} + (1-\nu)L_{C-Net} \tag{1}$$

$$L_{ID-Net} = -\sum_{t=1}^{T}[y\log(p_t) + (1-y)log(1-p_t)] \tag{2}$$

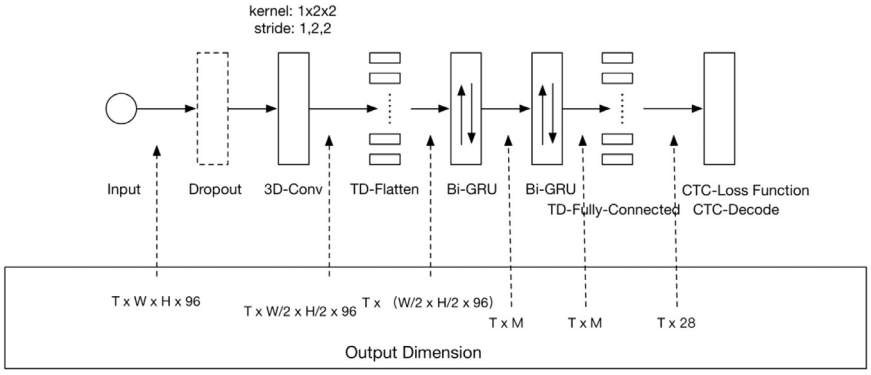

Fig. 7. The network structure of the C-Net.

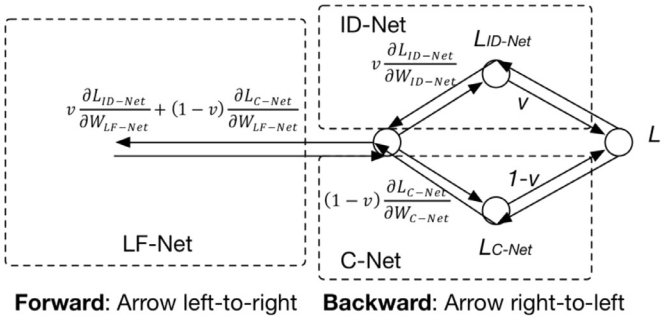

Fig. 8. Forward and backward propagation in Dual-Task Learning.

$$L_{C-Net} = -logp(l|x) \tag{3}$$

where v is a weighting factor balancing the importance of the identity-recognition and content-recognition tasks. L_{ID-Net} is the cross-entropy loss function of the ID-Net, where y is the actual class label (e.g. 1 for client and 0 for imposter) and $p_t, 1 \le t \le T$ is the output of the t-th neuron in the output layer. L_{C-Net} is the CTC loss function and l represents the label of the prompt text and x is the input sequence of the C-Net.

Denote the weights of the LF-Net, ID-Net, C-Net as $W_{LF\text{-}Net}$, $W_{ID\text{-}Net}$ and W_{C-Net}, respectively. The partial derivative of the loss function L with respect to the weights can be calculated as:

$$\frac{\partial L}{\partial W_{LF-Net}} = v\frac{\partial L_{ID-Net}}{\partial W_{LF-Net}} + (1-v)\frac{\partial L_{C-Net}}{\partial W_{LF-Net}} \tag{4}$$

$$\frac{\partial L}{\partial W_{ID-Net}} = v\frac{\partial L_{ID-Net}}{\partial W_{ID-Net}} \tag{5}$$

$$\frac{\partial L}{\partial W_{C-Net}} = (1-v)\frac{\partial L_{C-Net}}{\partial W_{C-Net}} \tag{6}$$

Using Eqn. (4) to Eqn. (6), all the weights can be optimized with any gradient-based nonlinear optimization algorithm such as Adam [34].

3.6. Implementation procedures

The implementation of the proposed LAN network includes four parts, i.e. the initialization stage, the training stage, the evaluation stage, and the test stage. Details of these stages are elaborated as follows.

(i) *Initialization Stage*: Since the number of training samples of each client is quite limited, a baseline network should be initialized before training the user-specific model. A world speaker set (denoted by S_w) containing N_w speakers is selected and for each speaker in S_w, $N_{T,w}$ prompt texts pronounced by the speaker will be used to train the baseline network (there are overall $N_w \times N_{T,w}$ training samples). Note that in the initialization stage, in order to help the ID-Net extract the identity-related features, the objective of the ID-Net is designed to recognize a speaker (a 1 against (N_w-1) classification) rather than to authenticate a speaker (a 1 to 1 classification). Hence, in the output layer of the ID-Net (as shown in Fig. 6), the number of neurons at each time spec is modified from 2 to N_w (each neuron represents a speaker label). Using the dual-tasks learning scheme, a baseline model M_{world} is obtained.

(ii) *Training Stage*: For each client, $N_{T,c}$ prompt texts pronounced by the speaker are employed as the client samples to obtain the user-specific model M_{client}. Note that in the output layer of the ID-Net in M_{client}, the number of neurons in the output layer at each time spec is set to 2 (as shown in Fig. 6, one for the client and the other for the imposter). Moreover, except for the weights linking to the neurons in the output layer of the ID-Net (which are initialized with the uniform distribution [37]), all the weights in M_{client} are initialized as those in M_{world}. Considering that the number of client samples (i.e. $N_{T,c}$) is much smaller than that of the imposter samples (i.e. $N_w \times N_{T,w}$), data augmentation by random shifting, zooming, flipping and rotation, and random sampling techniques is applied.

(iii) *Evaluation Stage*: For each client, $N_{E,c}$ prompt texts pronounced by the speaker and $N_{E,w}$ prompt texts pronounced by each speaker in the world speaker set S_w are employed as the client evaluation samples and the imposter evaluation samples, respectively. Note that the samples adopted in the evaluation stage should be different from those in the training and initialization stage. With the client model M_{client}, a threshold T_{client} is obtained by achieving a specific criterion (such as when equal error rate is obtained or when a very small false accept rate is obtained, etc.).

(iv) *Test Stage*: An unknown test sample is fed to the client model M_{client} corresponding to whom the speaker claimed to be. When the output of the model is greater than T_{client}, accept the request, or reject it otherwise.

Detailed implementation of the proposed LAN is given in the Appendix and we also provide the source code in https://github.com/klauscc/visual-speaker-authentication.

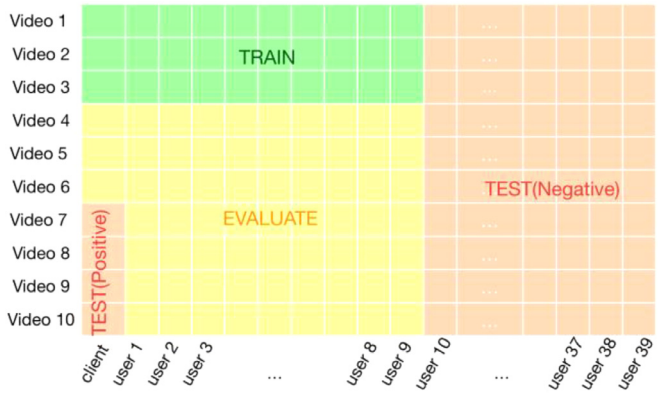

Fig. 9. The training, evaluation and test dataset partitioning.

4. Experiments and discussions

4.1. Discriminative power analysis on the ID-Net

To investigate the discriminative power of the proposed ID-Net in differentiating various speakers, a series of experiments on visual speaker authentication with a fixed password have been carried out. The visual speaker authentication dataset in [21] has been adopted for evaluation, which is composed of 40 speakers (29 male and 11 female) and each speaker was asked to pronounce the phrase "3725" in English (which is employed as the fixed password) for ten times. Each utterance contains ninety frames with a spatial resolution of 220×180.

Similar to [21], the Lausanne protocol [39] was employed in our experiment, which runs as follows and is illustrated in Fig. 9.

(i) *Training stage*: three clips of the client and thirty clips from ten other speakers are used as the positive and negative training samples, respectively. The client authentication model is then trained accordingly.
(ii) *Evaluation stage*: three clips of the client and the remaining seventy clips from the ten speakers are used as the positive and negative evaluation samples, respectively. The threshold T for authentication is tuned to obtain the equal error rate (*EER*), where the false accept rate (FAR_{eval}) equals to the false rejection rate (FRR_{eval}) in the evaluation set.
(iii) *Test stage*: the remaining four clips of the client and the 290 clips from the other 29 speakers are adopted as the client and imposter's test samples, respectively. The Half Total Error Rate (*HTER*) in the test set is computed as $HTER = (FAR_{test} + FRR_{test})/2$ by using the client model and the threshold T.
(iv) Finally, the average *EER* and *HTER* over the 40 speakers are computed to evaluate the authentication performance. Note that in order to avoid any bias in the selection of training samples, ten random trials have been performed and the average results are recorded.

In order to provide a comprehensive evaluation, three recent state-of-the-art visual speaker authentication approaches, i.e. Chan et al.'s approach (Chan's in short) [18], Liu et al.'s approach (Liu's in short) [20] and Lai et al.'s approach (Lai's in short) [21], are adopted for comparison. Among these approaches, Lai's, Chan's and the proposed ID-Net are region-based approaches and Liu's is a model-based approach which requires additional lip model information. Our previous proposed lip modelling approach [19] was employed to obtain the lip model and all the lip model results are manually examined to ensure that the obtained lip model is located correctly in the lip image.

It should be noted that in a practical visual speaker authentication system, the speaker-camera distance, the mouth position in the image and the angle of the mouth could vary greatly. In order to simulate such variations, we perform random affine transformations on the lip images sequences. Similar to [21], the translation (to simulate position variations), rotation (to simulate face angle variations) and scaling (to simulate speaker-camera distance variations) parameters range from −20 to 20 pixels (in both horizontal and vertical coordinates), −5°–5°, and 1–1.2, respectively.

Table 2 shows the authentication results using the proposed ID-Net and the other three approaches investigated. From the table, the following observations can be made: i) The model-based approach (Liu's) is more robust against the variations if the extracted lip model always locates on the accurate position; ii) The region-based approaches can achieve better authentication performance even without an accurate lip model. However, the authentication performance will be degraded to some extent when variations are introduced. Among all the three region-based approaches, the proposed ID-Net has the least performance degradation and the highest robustness. It is because the DCNN-based features can tolerate against variations caused by affine transformations; iii) The proposed ID-Net achieves the best (or comparable to the best) performance in both the authentication accuracy and the processing speed among all the approaches investigated and the accurate performance improvement is more obvious when variations are introduced.

4.2. Performance evaluation of the C-Net

The function of the C-Net is to authenticate whether the speaker has pronounced the prompt text, which can effectively protect the system from replay attacks. To solely evaluate the performance of the C-Net, the ID-Net is deactivated and only the LF-Net and C-Net are functional. The GRID database [36] is adopted for evaluation. It contains video recordings of 34 speakers (18 male, 16 female and the person with ID-21 was missing) and each speaker pronounced 1,000 sentences. Sentences are of the form "$command^{(4)} + color^{(4)} + preposition^{(4)} + letter^{(25)} + digit^{(10)} + adverb^{(4)}$", where the number in the bracket denotes the number of words in the category and there are 41 words in total. Specifically, each category consist of {*bin, lay, place, set*}, {*blue, green, red, white*}, {*at, by, in, with*}, {*A, …, Z*}\{*W*}, {*zero, …, nine*} and {*again, now, please, soon*}, respectively, yielding 64,000 possible sentences. For example, one sentence could be "*put red at G nine now*" .

As discussed in Section 3.6, considering that the number of client's training samples is quite limited, a world speaker set S_w containing N_w speakers is selected and all the utterances (1000 sequences in GRID) from each speaker in S_w are adopted to train the world model. Then for each client, $N_{E,c}$ prompt texts pronounced by the speaker are adopted to fine-tune the world model to generate the client model. During the test stage, the remaining utterances pronounced by the client are adopted as the test samples and the average character error rate (CER) between the actual prompt text and the recognized text by the C-Net is adopted as the performance evaluation metrics. CER is defined as the normalized edit distance of two sentences w.r.t character [29]. A small value of CER means the C-Net can successfully differentiate various uttering contents pronounced by the client and is highly robust against replay attacks, and vice versa. In Table 3, the average CERs achieved by the C-Net under different selections of N_w and $N_{E,c}$ are listed.

From Table 3, the following observations can be made. First, the character error rate is of a large value when N_w is relatively small (i.e. 5 or 10). It is mainly due to inadequate training samples. Specifically, there are about 64,000 possible prompt texts in total and the number of training samples are only 5,000 or 10,000. In such case, some words in the vocabulary have only been pro-

Table 2
Authentication performance comparisons using the four approaches investigated, where A/B denotes the results on the original database and the database with random affine transformations, respectively. The best results are highlighted. All the algorithms are run on a computer with two Intel Xeon Phi 2673v3 CPU (24 cores in total) and a graphic card of NVidia GeForce GTX 1080Ti.

		ID-Net	Lai's	Liu's	Chan's
EER_{eval} (%)		**0.00/0.00**	0.00/0.11	1.43/1.47	0.20/2.55
HTER (%)		0.60/**1.65**	**0.46**/2.34	4.59/4.75	1.56/8.00
Time (ms)	Feature Extraction	–	1631	113	2292
	Classification	–	90	13	121
	Total	**18**	1721	126	2413

Table 3
Performance evaluation in CER (%) by the C–Net with various selection of N_w and $N_{E,c}$.

No. of Speakers in the world set	No. of client training samples ($N_{E,c}$)			
N_w	25	50	100	200
5	38.1	36.3	33.3	31.6
10	17.2	15.4	12.5	11.8
20	5.7	5.3	4.7	3.6
30	4.2	3.8	2.9	2.5

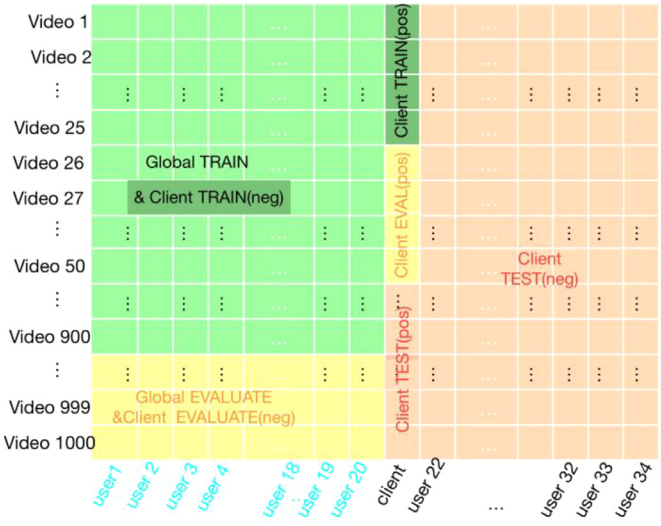

Fig. 10. The training, evaluation and test dataset partitioning in the GRID dataset.

nounced a few times, which cannot be well learned by the C-Net. Considering that there are overall 34 speakers in the GRID dataset, N_w is set to 20 as default for its low CER and relatively small number of training samples. Second, the CER reduces consistently as number of training samples from the client increases. It is because with more samples pronounced by the client, the C-Net can well grasp the talking habits and styles of the client and thus provide a high content authentication performance.

4.3. Visual speaker authentication with random prompt texts

In the random prompt texts scenario, the GRID database [36] is adopted for evaluation. To comprehensively evaluate the authentication performance, a random prompt text authentication protocol is designed based on the Lausanne protocol [39], which runs as follows and is illustrated in Fig. 10.

(i) *Initialization Stage*: Since each speaker only pronounces a limited number of prompt texts (1,000 out of 64,000 sequences) in the GRID dataset, twenty speakers ($N_w = 20$) are randomly selected from the dataset and 90% of their pronounced sequences (i.e. $N_{T,w} = 1000 \times 0.9 = 900$) were used to initialize the world-model, i.e. M_{world}.

(ii) *Training stage*: for each client, $N_{T,c}$ (which is set to 25 by default) utterance sequences are collected as client training samples and an overall $900 \times 20 = 18{,}000$ sequences in the initialization stage are used as the imposter samples. The client model M_{client} is then trained from M_{world} based on the above training samples.

(iii) *Evaluation Stage*: For each client, $N_{E,c}$ (which is set to 25 by default) sequences are used as the client evaluation samples. The remaining 2,000 sequences for the twenty speakers in the world speaker set S_w are used as the imposter evaluation samples (i.e. $N_{E,w} = 1000 \times 0.1 = 100$). With the client model M_{client}, the threshold T_{client} for authentication is tuned to the value when the false accept rate (FAR_{eval}) equals to a small value (0.01 in our experiment).

(iv) *Test Stage*: The remaining 1000-25-25 = 950 sequences for the client and all the sequences of the remaining twelve (33-20-1 = 12) unknown speakers are adopted as the test samples and the *HTER* is computed for each speaker. Similar to that in Section 4.1, ten random trials have been performed and the average values of *HTER* are employed to evaluate the authentication performance.

Table 4
Authentication results (in %) with different weight parameter selection.

v	*FRR*	FAR_c	FAR_{i1}	FAR_{i2}	*HTER*
1/3	10.04	2.46	0.93	0.03	5.59
1/2	7.03	2.54	0.66	0.02	4.05
2/3	**6.88**	2.55	**0.57**	**0.01**	**3.96**
3/4	11.64	**2.43**	0.89	0.02	6.38

4.3.1. Dual task learning analysis

As discussed in Section 3.5, the weight parameter v is designed to balance the network emphasis between the identity and content discrimination tasks. When v is small, the proposed LAN focused more on the content information and thus the features extracted by the LF-Net is more related to the linguistic information. On the other hand, the output of the LF-Net is more related to the identity information when v is large. To achieve optimal performance, several typical selections of v are investigated and the corresponding authentication results are listed in Table 4. Note that to further investigate the authentication error, the false accept rate (FAR) is decomposed into the following three parts, i.e. FAR_c representing the rate of the client pronouncing the incorrect prompt text, FAR_{i1} and FAR_{i2} representing the rate of the imposter pronouncing the correct and incorrect text, respectively.

From Table 4, the following observations can be made. First, when v is 1/2 or 2/3, the proposed network can achieve a relatively low *FRR* with acceptable *FARs* (FAR_c around 2.5%, $FAR_{i1} < 1\%$ and FAR_{i2} close to 0%). Moreover, it should be noted that when

Table 5
Authentication results (in %) for various clients.

Speaker	*FRR*	FAR_c	FAR_{i1}	FAR_{i2}	*HTER*
22	8.95	2.47	1.75	0.04	5.19
23	1.37	2.68	0.22	0.01	1.17
24	8.00	2.47	0.00	0.00	4.41
25	5.26	2.51	0.32	0.01	3.10
26	4.58	2.43	0.06	0.00	2.71
27	5.79	2.69	0.14	0.00	3.37
28	7.85	2.36	0.00	0.00	4.32
29	5.69	2.54	3.28	0.09	3.83
30	10.63	2.45	0.00	0.00	5.72
31	3.79	2.57	0.11	0.00	2.34
32	9.37	2.45	0.00	0.00	5.09
33	3.37	2.61	0.42	0.01	2.19
34	13.47	2.48	0.00	0.00	7.15
Average	**6.78**	**2.52**	**0.48**	**0.01**	**3.89**

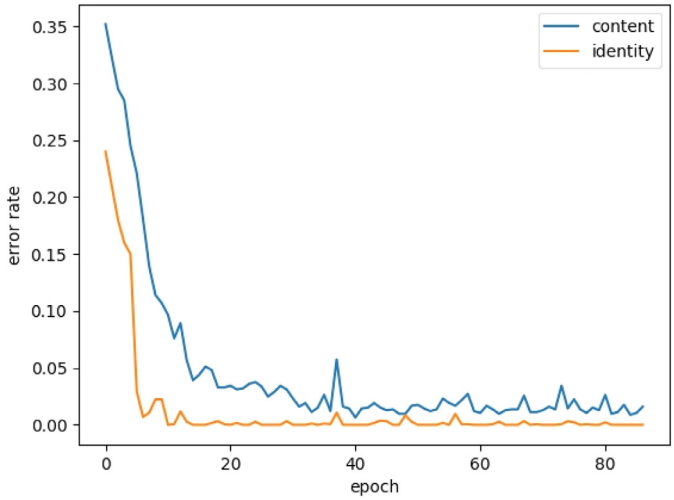

Fig. 11. Error curves in dual-task training.

v is around [1/2, 2/3], the authentication performance does not change much, which demonstrates that our approach is not sensitive to the exact value of v. Second, when v is too big (e.g. 3/4) or too small (e.g. 1/3), both the *FRR* and FAR_{i1} will increase, which demonstrates that the two tasks are complementary and focusing too much on either task will degrade the performance of the authentication system. Finally, the optimal v is set to 2/3, which achieves the lowest *HTER*.

Fig. 11 shows the identity and content error curves during the training procedure of M_{world}. Note that in each epoch, 80% of the samples in the training set are used to train the network and the remaining samples are adopted as validation samples to compute the identity authentication and content recognition error. From the figure, it is observed that during dual-task learning, both the identity and content errors are being reduced gradually, which demonstrates the effectiveness of the proposed dual-task training scheme. The features extracted by the LF-Net can serve both the identity authentication and the content recognition tasks. Comparatively speaking, the identity authentication error rate has been optimized to a small value (e.g. 0.01) in around 20 epochs. On the other hand, the content recognition errors converge more slowly and the converged content recognition error rate is greater than that of the identity authentication error rate. It is because some words in the vocabulary in the GRID database are not visually differentiable. The detailed error analysis is given in the following subsection.

4.3.2. Error analysis of the authentication results

Compared with the fixed password scenario, speaker authentication with random prompt texts has a relatively higher *HTER* (3.96% against 1.65%). In order to analyze the key factors leading to the authentication error, the following experiments have been carried out. Without losing generality, the No.1 to No.20 speakers in the GRID database are used to construct the world speaker set S_w and the remaining speakers (the No.22 to No.34, No.21 speaker is missing in GRID) are used to construct the client speaker set. The authentication results are given in Table 5.

From Table 5, it is observed that for most speakers, the system achieves a very low *FAR* for imposters even if they pronounce the correct prompt text; however, it is more likely for the system to erroneously accept a client sequence pronouncing the incorrect prompt text. It is because in the GRID database, some words (e.g. “b” and “p”) in the visual domain are difficult to differentiate, which leads to the false acceptance. Moreover, each speaker is asked to pronounce 1,000 random sequences in the GRID database and FAR_c will be relatively larger if he/she pronounces more visually similar prompt texts.

Fig. 12 shows the confusion matrices for various contents and identities. It is observed from the figure that some sets in the prompt texts are of higher discriminative power (i.e. the “command”, “color”, “digit” and “adverb”) than the rest (i.e. “preposition” and “letter”). Especially in the “letter” set, it usually takes a short period to pronounce a letter and many letters are visually similar to each other (e.g. “b” and “p”, “d” and “t”, etc.), which brings great difficulties for content authentication. It is suggested that in order to guarantee a high authentication performance, the words in the prompt texts should be as visually distinctive as possible.

4.3.3. Authentication results with various numbers of training samples

As we known, the client will be impatient if he/she is asked to pronounce specific texts for too many times during the training stage. To investigate the relationship between the authentication performance and the number of training samples, an experiment has been carried out and the results are given in Table 6. Note that in the experiment, the client model trained in Section 4.3.1 (with 25 training and 25 evaluation client samples) is adopted as the baseline model. When the number of training samples increases, the new model is obtained by fine-tuning the baseline model with the new samples.

From the table, it is observed that *HTER* decreases with the increase of the number of training samples. As the threshold is selected with a fixed FAR_{eval} (0.01), the performance improvement is reflected in the decrease of *FRR*. Hence, the proposed authentica-

Table 6
Authentication results with various client training samples. We randomly select 25, 50, 100, 200 samples for training, 25 samples for evaluation, and the remaining samples for testing.

No. of client samples for training & evaluation	*FRR*	FAR_c	FAR_{i1}	FAR_{i2}	*HTER*
50 (25 + 25)	6.59	2.51	0.79	0.02	3.85
75 (50 + 25)	5.61	2.55	0.63	0.02	3.34
125 (100 + 25)	3.33	2.60	0.93	0.03	2.25
225 (200 + 25)	2.97	2.61	0.66	0.02	**2.03**

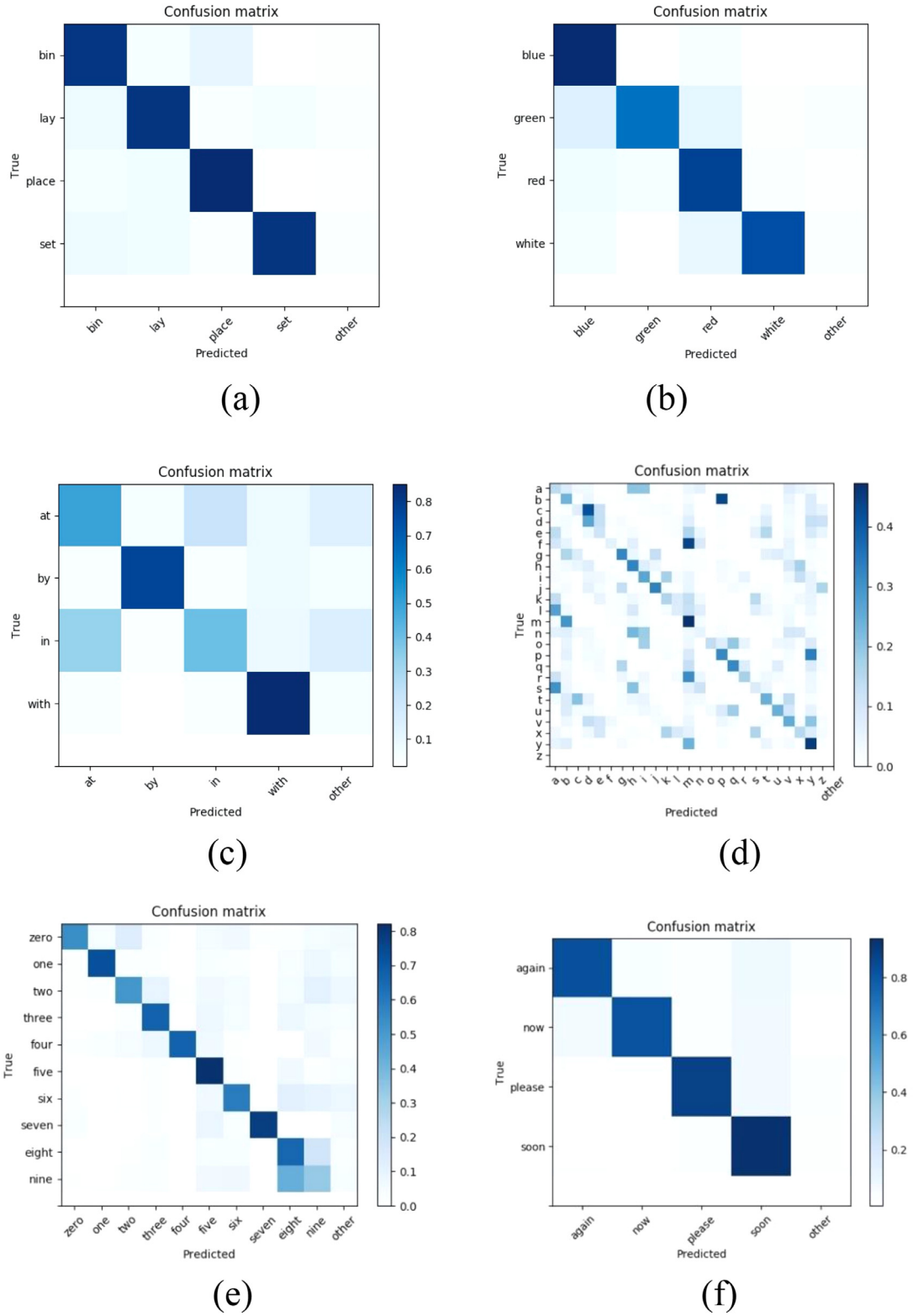

Fig. 12. The confusion matrices for (a)–(f): various words in the prompt text.

tion system can start with a reliable baseline model using limited client training samples (50 sequences for a prompt text vocabulary with 41 words) and the authentication performance will continue to increase by updating the baseline model with new client samples (which are obtained from authorized sequences during authentication).

5. Conclusion

In this paper, a new DCNN-based approach is proposed for visual speaker authentication with random prompt texts. The face image sequence of a speaker pronouncing a specific prompt text is captured and the Dlib face detector is adopted to extract the lip

Table A1
LF-Net structure. The total number of parameters is about 1.00 million.

Layer name	Input size	Kernel/Stride/Pad/No. of Filters	Parameter #
Conv3D_1	$T\times 50\times 100\times 3$	$3\times 5\times 5$/ 1, 2, 2/ 1, 2, 2/ 32	7232
3D-Res-block_1	$T\times 25\times 50\times 32$	–/–/–/ 32	55,360
Conv3D_2	$T\times 25\times 50\times 32$	$3\times 3\times 3$/ 1, 2, 2/ 1, 1, 1/ 64	55,360
3D-Res-block_2	$T\times 13\times 25\times 64$	–/–/–/ 64	221,312
Conv3D_3	$T\times 13\times 25\times 64$	$3\times 3\times 3$/ 1, 2, 2/ 1, 1, 1/ 96	165,984
3D-Res-block_3	$T\times 7\times 13\times 96$	–/–/–/ 96	497,856

Table A2
ID-Net structure. The total number of parameters is about 1.22 million.

Layer name	Input size	Kernel/Stride/Pad/No. of Filters (or Parameter)	Parameter #
TD-Conv2D_1	$T\times 7\times 13\times 96$	3×3/ 2, 2/ 1, 1/ 128	110,592
TD-2D-Res-block_1	$T\times 4\times 7\times 128$	–/–/–/ 256	589,842
TD-Ave-Pooling_1	$T\times 4\times 7\times 256$	(2, 3)/–/–/–	0
Dropout	$T\times(2\times 2\times 256)$	$p=0.5$	0
TD-FC_1	$T\times 1024$	512	524,288
Dropout	$T\times 512$	$p=0.5$	0
TD-FC_2	$T\times 512$	No. of outputs = 2	1024

Table A3
C–Net structure. The total number of parameters is about 5.74 million.

Layer name	Input size	Kernel/Stride/Pad/No. of Filters (or Parameter)	Parameter #
SpatialDropout3D_1	$T\times 7\times 13\times 96$	$p=0.5$	0
Conv3D_4	$T\times 7\times 13\times 96$	1 * 2 * 2/ 1, 2, 2/ 0,1,1/ 96	36,864
Bi_GRU_1	$T\times 4\times 7\times 96$	512/–/–/–	4,523,520
Bi_GRU_2	$T\times 512$	512/–/–/–	1,181,184
TD-FC_3	$T\times 512$	No. of outputs = 28	14,364

region. Then a lip feature network, i.e. the LF-Net, is designed to extract representative lip features from the lip image sequences, which can comprehensively describe the static and dynamic information of the lip region. To differentiate different speakers and the articulated contents, an identity network, i.e. the ID-Net, and a content network, i.e. the C-Net, have been designed, respectively. A dual-tasks learning scheme is employed to train the above three networks simultaneously and thus the trained network can authenticate both the speaker's identity and the speech content. Experimental results have demonstrated that the proposed ID-Net can achieve high authentication performance and is also robust against variations caused by various talking poses or distances towards the camera. Compared with several state-of-the-art visual speaker authentication approaches in the fixed-password scenario, the proposed ID-Net achieves the best performance. Moreover, we show that in the visual speaker authentication scenario with random prompt texts, the proposed approach can achieve high authentication accuracies and the authentication performance can be further improved with more training samples. To the best of our knowledge, there are very few works in the literature considering the random prompt texts scenario and our approach is the first to provide a feasible solution in this area. Considering that multimodal learning and feature fusion techniques have shown great advantages in many AI based systems [41,42], we plan to extend the proposed scheme by incorporating it with other face and/or voice based authentication systems in our future work to achieve a higher level of security.

Acknowledgement

The work described in this paper is fully supported by NSFC Fund (No. 61771310).

Appendix.

The proposed LAN is implemented using Keras with a Tensorflow backend. The detailed structure of LF-Net, ID-Net, C-Net is in Table A1–A3 . Note that a Batch Normalization [37] and a ReLU [38] activation layer follow each convolution layer, which is not listed in the tables. T denotes the number of frames in each lip image sequence and it is set to 75 for GRID dataset [36] and 90 for the authentication dataset in [21], respectively. The network parameters are initialized with the normal initializer [40]. The loss function of ID-Net and C-Net is the cross-entropy error function and CTC loss function [32], respectively. The proposed LAN is trained end-to-end by Dual-Task Learning described in Section 3.5 with the loss weight parameter of 2/3. In the optimization stage, an auto-adjust-learning rate gradient-based algorithm, i.e. Adam [34], is employed with a learning rate of 0.001 and the default hyperparameters are set as follows: two momentum coefficients as 0.9 and 0.999, fuzz factor as 10^{-8} and no learning rate decay.

A world-model M_{world} is trained in the initialization stage where the final output of ID-Net is $T\times N_w$, where N_w is the number of speakers. For each client, a client model M_{client} is transferred from M_{world}, i.e. all the weight parameters of M_{client} are initialized with those of M_{world}, except the final layer of ID-Net whose output is changed from $T\times N_w$ to $T\times 2$.

References

[1] D. Gragnaniello, G. Poggi, C. Sansone, L. Verdoliva, Local contrast phase descriptor for fingerprint liveness detection, Pattern Recognit. 48 (4) (2015) 1050–1058.
[2] P. Wild, P. Radu, L. Chen, J. Ferryman, Robust multimodal face and fingerprint fusion in the presence of spoofing attacks, Pattern Recognit. 50 (2016) 17–25.
[3] R. Frischholz, U. Dieckmann, Biold: a multimodal biometric identification system, Computer 33 (2) (2000) 64–68.
[4] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp. 770–778.

[5] G. Huang, Z. Liu, K.Q. Weinberger, L. van derMaaten, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp. 2261–2269.
[6] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, IEEE Trans. Image Process. 24 (12) (2015) 5659–5670.
[7] C. Hong, J. Yu, D. Tao, M. Wang, Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval, IEEE Trans. Indust. Electron. 62 (6) (2015) 3742–3751.
[8] M. Patacchiola, A. Cangelosi, Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods, Pattern Recognit. 71 (2017) 132–143.
[9] J.C. Núñez, R. Cabido, J.J. Pantrigo, A.S. Montemayor, J.F. Vélez, Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition, Pattern Recognit. 76 (2018) 80–94.
[10] T Wark, D Thambiratnam, S Sridharan, Person authentication using lip information, in: Proceedings of TENCON'97. IEEE region 10 annual conference. speech and image technologies for computing and telecommunications, 1, 1997, pp. 153–156.
[11] P. Jourlin, J. Luettin, D. Genoud, H. Wassner, Acoustic-labial speaker verification, Pattern Recognit. Lett. 18 (9) (1997) 853–858.
[12] C.C. Broun, X. Zhang, R.M. Mersereau, M. Clements, Automatic speechreading with application to speaker verification, in: In proceedings of 2002 IEEE international conference on acoustics, speech, and signal processing (ICASSP), 1, 2002, pp. 685–688.
[13] H. Ertan Cetingul, Yücel Yemez, Engin Erzin, A. Murat Tekalp, Discriminative analysis of lip motion features for speaker identification and speech-reading, IEEE Trans. Image Process. 15 (10) (2006) 2879–2891.
[14] M.I. Faraj, J. Bigun, Motion features from lip movement for person authentication, in: In proceedings of 18th international conference on pattern recognition (ICPR), 3, 2006, pp. 1059–1062.
[15] U.R. Sanchez, J. Kittler, Fusion of talking face biometric modalities for personal identity verification, in: In proceedings of 2006 IEEE International conference on acoustics, speech, and signal processing (ICASSP), 5, 2006, pp. 1073–1076.
[16] S.A. Samad, D.A. Ramli, A. Hussain, Lower face verification centered on lips using correlation filters, Inf. Technol. J. 6 (8) (2007) 1146–1151.
[17] Y.F. Liu, C.Y. Lin, J.M. Guo, Impact of the lips for biometrics, IEEE Trans. Image Process. 21 (6) (2012) 3092–3101.
[18] C.H. Chan, B. Goswami, J. Kittler, W. Christmas, Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication, IEEE Trans. Inf. Forensics Secur. 7 (2) (2012) 602–612.
[19] S.L. Wang, A.W.C. Liew, Physiological and behavioral lip biometrics: a comprehensive study of their discriminative power, Pattern Recognit. 45 (9) (2012) 3328–3335.
[20] X. Liu, Y.M. Cheung, Learning multi-boosted hmms for lip-password based speaker verification, IEEE Trans. Inf. Forensics Secur. 9 (2) (2014) 233–246.
[21] J.Y. Lai, S.L. Wang, A.W.C. Liew, X.J. Shi, Visual speaker identification and authentication by joint spatiotemporal sparse coding and hierarchical pooling, Inf. Sci. 373 (2016) 219–232.
[22] K Suzuki, Y Tsuchihashi, H Suzuki, A trail of personal identification by means of lip print, Jpn. Soc. Legal Med. 22 (1968) 392.
[23] J.R. Movellan, Visual speech recognition with stochastic networks, in: Proceedings of advances in neural information processing systems, 1995, pp. 851–858.
[24] S Pigeon. The m2vts database, laboratoire de telecommunications et teledection, place du levant. 1996.
[25] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, Xm2vtsdb: the extended m2vts database, in: In proceedings of second international conference on audio and video-based biometric person authentication, 964, 1999, pp. 965–966.
[26] C.W. Liao, W.Y. Lin, C.W. Lin, Video-based person authetication with random passwords, in: In proceedings of 2008 IEEE international conference on multimedia and expo, 2008, pp. 581–584.
[27] V. Kazemi, S. Josephine, One millisecond face alignment with an ensemble of regression trees, in: In proceedings of 2014 IEEE conference on computer vision and pattern recognition (CVPR), 2014, pp. 1867–1874.
[28] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K.J. Lang, Phoneme recognition using time-delay neural networks, IEEE Trans. Acoust. Speech Signal Process. 37 (1989) 328–339.
[29] Y.M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. Lipnet: Sentence-level lipreading, arXiv:1611.01599, 2016.
[30] J. Chung, C. Gulcehre, K.H. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555, 2014.
[31] F.A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: continual prediction with LSTM, Neural Comput. 12 (10) (2000) 2451–2471.
[32] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd international conference on machine learning (ICML), 2006, pp. 369–376.
[33] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: Proceedings of the 25th international conference on machine learning (ICML), 2008, pp. 160–167.
[34] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.
[35] X.X. Shi, S.L. Wang, J.Y. Lai, Visual speaker authentication by ensemble learning over static and dynamic lip details, in: Proceedings of 2016 IEEE international conference on image processing (ICIP), 2016, pp. 3942–3946.
[36] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, J. Acoust. Soc. Am. 120 (5) (2006) 2421–2424.
[37] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167, 2015.
[38] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th international conference on machine learning (ICML), 2010, pp. 807–814.
[39] J. Luettin, G. Maitre, Evaluation protocol for the extended M2VTS database (XM2VTSDB), IDIAP communication, 1998.
[40] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision (ICCV), 2015, pp. 1026–1034.
[41] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: In 2017 IEEE international conference on computer vision (ICCV), 3, 2017, pp. 1839–1848.
[42] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao. Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. To appear in IEEE transactions on neural networks and learning systems.

Feng Cheng received his B.Eng. degree in Information Security from Shanghai Jiaotong University, Shanghai, China in 2016. Since 2016, he has been pursuing his Master's degree in the School of Cyber Security, Shanghai Jiaotong University. His research interests include computer vision and pattern recognition.

Shi-Lin Wang received his B.Eng. degree in Electrical and Electronic Engineering from Shanghai Jiaotong University, Shanghai, China in 2001, and his Ph.D. degree in the Department of Computer Engineering and Information Technology, City University of Hong Kong in 2004. Since 2004, he has been with the School of Information Security Engineering, Shanghai Jiaotong University, where he is currently an Associate Professor. His research interests include image processing and pattern recognition. Dr. Wang is a senior member of the Institute of Electrical and Electronic Engineers (IEEE) and his biography is listed in Marquis Who's Who in Science and Engineering.

Alan Wee-Chung Liew received his B.Eng. with first class honors in Electrical and Electronic Engineering from the University of Auckland, New Zealand, in 1993, and Ph.D. in Electronic Engineering from the University of Tasmania, Australia, in 1997. He worked as a Research Fellow and later a Senior Research Fellow at the Department of Electronic Engineering at the City University of Hong Kong. From 2004 to 2007, he was with the Department of Computer Science and Engineering, The Chinese University of Hong Kong as an Assistant Professor. In 2007, he joined the School of Information and Communication Technology, Griffith University as an Associate Professor. His current research interests include computer vision, medical imaging, pattern recognition and bioinformatics. He serves as a technical reviewer for many international conferences and journals such as IEEE Transactions, IEE proceedings, bioinformatics and computational biology. Dr. Liew is a senior member of the Institute of Electrical and Electronic Engineers (IEEE) since 2005, and his biography is listed in the Marquis Who's Who in the World and Marquis Who's Who in Science and Engineering.