Efficient Audio-Visual Speaker Recognition via Deep Heterogeneous Feature Fusion

Yu-Hang Liu[1,2], Xin Liu[1,2](✉), Wentao Fan[1,2],
Bineng Zhong[1,2], and Ji-Xiang Du[1,2]

[1] Department of Computer Science, Huaqiao University, Xiamen 361021, China
xliu@hqu.edu.cn
[2] Xiamen Key Laboratory of Computer Vision and Pattern Recognition,
Huaqiao University, Xiamen 361021, China

Abstract. Audio-visual speaker recognition (AVSR) has long been an active research area primarily due to its complementary information for reliable access control in biometric system, and it is a challenging problem mainly attributes to its multimodal nature. In this paper, we present an efficient audio-visual speaker recognition approach via deep heterogeneous feature fusion. First, we exploit a dual-branch deep convolutional neural networks (CNN) learning framework to extract and fuse the high-level semantic features of face and audio data. Further, by considering the temporal dependency of audio-visual data, we embed the fused features into a bidirectional Long Short-Term Memory (LSTM) networks to produce the recognition result, though which the speakers acquired under different challenging conditions can be well identified. The experimental results have demonstrated the efficiency of our proposed approach in both audio-visual feature fusion and speaker recognition.

Keywords: Audio-visual speaker recognition · Deep heterogeneous feature fusion · Dual-branch deep CNN · Bidirectional LSTM

1 Introduction

Multi-modal biometric person recognition has received a lot of attention in recent years due to the growing security demands in commercial and law enforcement applications. In particular, speaker recognition is one of the active research problems in biometric community, and audio-visual (AV) biometrics generally offer complementary information sources for speaker identity characterization. Among them, face and voice features, incorporating the advantages of non-intrusiveness and easy acquisitions, have become economically feasible, but the appropriate fusion between these two heterogeneous modalities is still a non-trivial task.

In the past, different kinds of approaches have been exploited to fuse the face and voice data. In general, the audio-visual integration can be divided into four categories: sensor-level, feature-level, matching-level and decision-level. Since the sensor-level based fusion approaches require that the input data types must be

J. Zhou et al. (Eds.): CCBR 2017, LNCS 10568, pp. 575–583, 2017.
https://doi.org/10.1007/978-3-319-69923-3_62

the same, such that there are more matching-level and decision-level fusions. For instance, Cheng et al. [2] utilized the proposed IKFD method to obtain the face recognition scores and employed GMMs to produce the voice recognition scores then fused the scores. Similarly, Feng et al. [3] utilized the GMM to fuse face scores and audio scores. Later, Soltane et al. [12] addressed an adaptive Bayesian method to fuse the scores of face and speech modalities. These matching-level or decision-level fusion functions could not take full advantage of the information.

To utilize more information, some researchers attempted to fuse the face and voice in feature level module. In the early years, researchers mainly used common feature transformation functions. For instance, Bredin et al. [1] utilized the canonical correlation analysis (CCA) to fuse the audio-visual features for speaker recognition, while Haghighat et al. [6] proposed discriminant correlation analysis (DCA) to fuse the audio-visual features for identification. These methods would cause information loss when transform features.

Recently, deep networks have been successfully applied to unsupervised feature learning for multimodal deep learning [10]. Benefit from this finding, Hu et al. [7] and Geng et al. [4] used CNN to fuse face features and audio features and achieved good results. However, they did not find the position in CNN that is most suitable for feature fusion. Ren et al. presented a multimodal LSTM networks for speaker identification, but the features are not fused actually.

In this paper, as shown in Fig. 1, we present an efficient audio-visual speaker recognition approach via deep heterogeneous feature fusion. The proposed approach first exploits a dual-branch deep CNN learning framework to extract the high-level semantic features of face and audio data, whereby the learned heterogeneous features between these two modalities can be well fused. Further, by considering the temporal dependency of audio-visual data, we embed the fused features into a bidirectional Long Short-Term Memory (LSTM) network to produce the speaker recognition result, featuring more discriminative power. The experimental results have its outstanding performance.

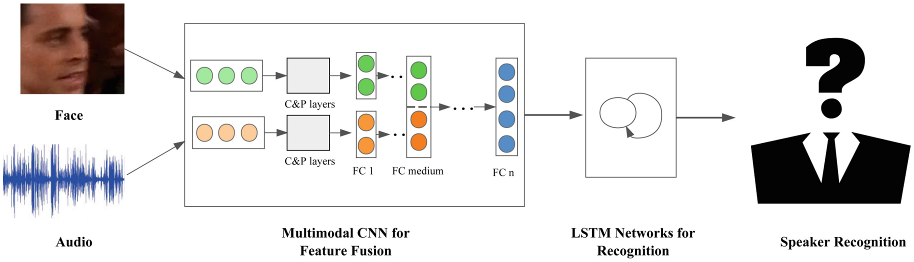

Fig. 1. The pipline of our proposed speaker recognition framework, in which the face features and audio features are fused by our proposed dual-branch deep CNN model.

2 Feature Fusion and Recognition Architecture

In this part, we explore how to use CNN to extract and fuse the features of face and audio, and further propose a dual-branch CNN model for feature extraction and fusion. In addition, we utilize the bidirectional LSTM networks associated with the fused information to get a reliable recognition result.

2.1 Dual-Branch CNN Model for Feature Fusion

In our deep feature fusion learning architecture, face features and audio features are extracted via the CNN model, which consists of convolutional and pooling layers and fully connected layers:

$$h_i = \begin{cases} P(\sigma(conv(W_i, h_{i-1}) + b_i)), \; i = 1, \ldots, m, \\ \sigma(W_i \cdot h_{i-1} + b_i), \; i = m+1, \ldots, n, \end{cases} \tag{1}$$

where h_i is the output of the i-th layer and h_0 is the raw input of the networks, W_i is the weight matrix and b_i is the bias term for the i-th layer, σ stands for the nonlinear activation function, e.g., tanh, sigmoid, or ReLU [9], P represents the pooling function. To explore the best position for fusion in CNN, four kinds of dual-branch CNN models are illustrated as shown in Fig. 2.

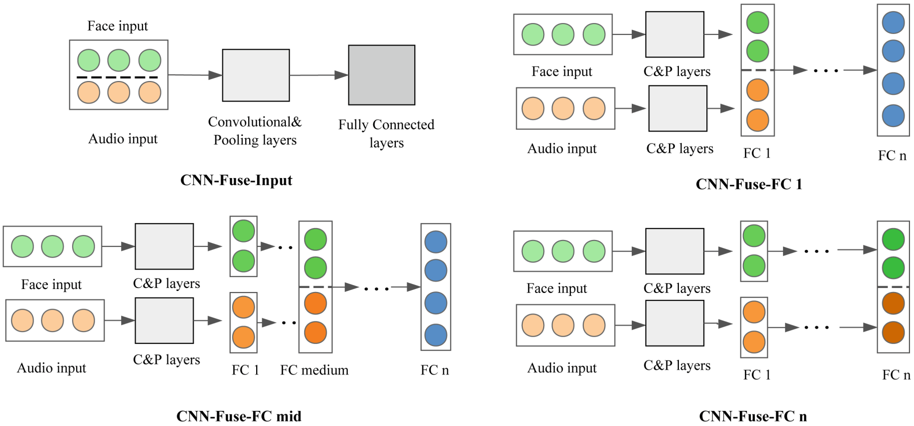

Fig. 2. Four kinds of dual-branch CNN Models for feature fusion.

CNN-Fuse-Input: In this case, the raw face input and audio input are concatenated as the input of CNN:

$$h_0 = concatenate(h_0^{face}, h_0^{audio}) \tag{2}$$

CNN-Fuse-FC 1: Convolved face features and audio features are concatenated as the input of fully connected layers:

$$h_i^j = P(\sigma(conv(W_i^j, h_{i-1}^j) + b_i^j)),\ i = 1, \ldots, m, \tag{3}$$

$$h_m = concatenate(h_m^{face}, h_m^{audio}), \tag{4}$$

where j represents the modality, i.e. face or audio.

CNN-Fuse-FC mid: Face and audio features extracted by different C&P layers and FC layers are concatenated as the input of the remaining FC layers:

$$h_i^j = \begin{cases} P(\sigma(conv(W_i^j, h_{i-1}^j) + b_i^j)),\ i = 1, \ldots, m, \\ \sigma(W_i^j \cdot h_{i-1}^j + b_i^j),\ i = m+1, \ldots, m+mid, \end{cases} \tag{5}$$

$$h_{m+mid} = concatenate(h_{m+mid}^{face}, h_{m+mid}^{audio}), \tag{6}$$

CNN-Fuse-FC n: We first use two CNN models to extract deep face features and audio features separately, and then concatenate them:

$$h_n = concatenate(h_n^{face}, h_n^{audio}), \tag{7}$$

After training, the deep fused features could be extracted from h_n directly.

2.2 Bidirectional LSTM Networks for Recognition

In general, the face images are always influenced by the bad image quality, exaggerated expression, or illumination, which would degrade the recognition accuracy. To solve this problem, the bidirectional LSTM networks incorporating the temporal modeling ability is employed. In LSTM networks the hidden units are LSTM cells. The spirit of the LSTM cell is that for every step the cell would choose some information to "remember" and some to "forget", so that LSTM networks could learn longer information dependencies than simple RNN. In particular, bidirectional LSTM networks has been proved to be more effective for recognition [5]. Therefore, BILSTM is employed for recognition purpose. Specifically, three different methods are employed to get the recognition result:

$$y'_{last} = softmax(h_n) \tag{8}$$

$$y'_{vote} = \max_i(\sum_{t=1}^{n} I(y'_t = i)), i = 1, \ldots, c \tag{9}$$

$$y'_{mean} = softmax(W \cdot (\frac{1}{n}\sum_{t=1}^{n} h_t)) \tag{10}$$

In Eq. (8), we select the last output as the final result. In Eq. (9), the final result is generated by voting of the outputs of every results. In Eq. (10), we average the outputs of every step and utilize the softmax to classify the average value. Our whole feature fusion and recognition framework can be expressed in Fig. 3.

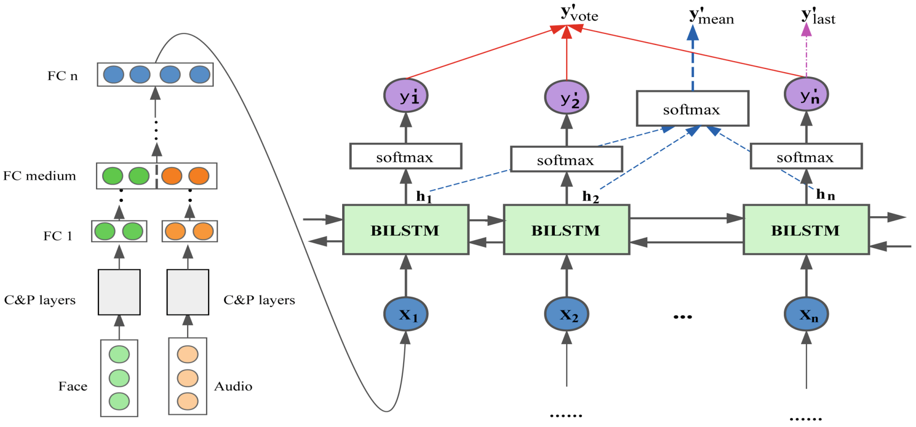

Fig. 3. Our proposed feature fusion and recognition framework.

3 Experiments and Results

The public available audio-visual dataset collected by Hu et al. [7] are selected for the evaluation. The face images and audio clips are extracted from nine episodes of two TV series, i.e. "Friends" and "The Big Bang Theory"("BBT"), in which the leading roles are selected for testing, including six actors in "Friends", i.e., Rachel, Monica, Phoebe, Joey, Chandler and Ross, and five actors in "BBT", i.e., Sheldon, Leonard, Howard, Raj and Penny. For "Friends", the faces and audio are collected from five episodes of different seasons, and we use the data of S01E03 (Season 01, Episode 03), S04E04, S07E07 and S10E15 for training and the data of S05E05 for testing. In total, there are 87273 face images involved for training and 29539 face images for testing. For "BBT", we choose S01E04, S01E05, S01E06 for training and S01E03 for testing, and total numbers of faces for training and testing are 90034 and 28554, respectively.

3.1 Multimodal CNN Model for Feature Fusion

For feature fusion, we first resize all the face images to 50×50 and convert them to gray-level images, and dimension of each face vector is 2500. Similar to the work [7], we utilize the mel frequency cepstral coefficients (MFCCs) [11] to preliminarily extract audio feature, and we acquire a 375D feature vector for every audio sample.

After extracting the primary features, we carried experiments on the four different feature fusion models, all the configurations in CNN are the same in different models. In addition, we add dropout [13] and batch normalization [8] to optimize our networks. Table 1 shows the recognition accuracies on "Friends" dataset and "BBT" dataset of different feature fusion models.

It can be found that the model "CNN-Fuse-Input" performed even worse than face modality only, which indicates that the raw features of different modalities

Table 1. Recognition accuracy of different feature fusion models

Model	Accuracy(%) on “Friends”	Accuracy(%) on “BBT”
Only face	94.0	93.6
CNN-Fuse Input	92.3	92.1
CNN-Fuse-FC 1	94.6	95.0
CNN-Fuse-FC mid	**95.6**	**95.4**
CNN-Fuse-FC n	95.0	94.9

are not suitable for fusion directly. Meanwhile, the “CNN-Fuse-FC mid” model has produced a better result than that of “CNN-Fuse-FC 1” model. That is, the high-level features extracted by CNN model are suitable for fusion. Note that, the “CNN-Fuse-FC mid” model also performs better than model “CNN-Fuse-FC n”, that is because in the last two fully connected layers the concatenated features are fused better by nonlinear feature transformation. We can conclude that the middle layer of fully connected layers is the best place for feature fusion.

Table 2. Recognition accuracy of different feature fusion methods.

Method	Accuracy(%) on “Friends”	Accuracy(%) on “BBT”
PCA+LDA+SVM	84.0	85.4
PCA+MDA+SVM	83.0	83.9
CCA+SVM	82.9	83.4
DCA+SVM	84.5	85.6
Hu et al.	88.5	-
CNN-Fuse-FC mid	**95.6**	**95.4**

In order to prove the effectiveness of our model for feature fusion, we also conducted experiments of some common feature fusion methods mentioned above and contrasted the experimental results, in which the typical SVM [14] was chosen to classify the fused features. The recognition results obtained by different approaches were listed in Table 2. It can be found that our feature fusion model is more effective than the common feature fusion methods. The main reason lies that the deep learning networks has an advantage in automatic feature transformation and extraction.

3.2 Bidirectional LSTM Networks for Recognition

For speaker recognition, we select to extract the fused features from the last fully connected layer as the input of bidirectional LSTM networks, in which the dimension of fused face and its corresponding audio vector is 1000. There are 24 frames per second in the video, and we choose 0.5 s for temporal characterization. In the experiments, the fused features extract from the aforementioned four models are embedded into the bidirectional LSTM networks for recognition. Meanwhile, we compare the proposed approach with the basic voting scheme, and the recognition results are shown in Table 3.

Table 3. Recognition accuracy obtained by different fusions and classifiers.

Method	Accuracy(%) on “Friends”	Accuracy(%) on “BBT”
Only face+BILSTM	96.4	96.1
CNN-Fuse-Input+BILSTM	94.5	94.0
CNN-Fuse-FC 1+BILSTM	97.4	97.4
CNN-Fuse-FC n+BILSTM	97.6	97.2
CNN-Fuse-FC mid+vote	97.3	97.2
CNN-Fuse-FC mid+LSTM	97.8	97.5
CNN-Fuse-FC mid + BILSTM	**98.2**	**97.8**

It can be clearly observed that the “CNN-Fuse-FC mid” model associated with the bidirectional LSTM networks has achieved the best results. That is, “CNN-Fuse-FC mid” is more discriminative for audio-visual heterogeneous feature fusion. Under the same fused features, the bidirectional LSTM networks have produced the better result than ordinary LSTM networks and voting scheme. As shown in Fig. 4, we also implemented three methods mentioned above to get the recognition result of bidirectional LSTM networks. From the experimental results, it can be found that the voting scheme and the utilization of mean value perform nearly and better than selection of last step.

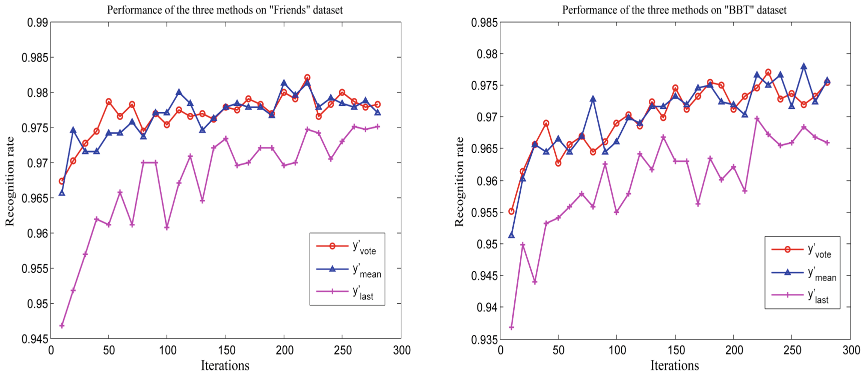

Fig. 4. Performance of the three operations in bidirectional LSTM networks.

4 Conclusion

In this paper, we have presented an efficient audio-visual speaker recognition approach via deep heterogeneous feature fusion. The proposed approach exploits a dual-branch deep CNN learning framework to extract and fuse the face and voice features in high-level semantic space. Meanwhile, by considering the temporal dependency of audio-visual fused features, a bidirectional Long Short-Term Memory networks is utilized to produce the recognition result. Accordingly, the speakers acquired under different challenging conditions can be well identified. The experimental results have shown that our proposed audio-visual speaker recognition approach performs well in both feature fusion and speaker recognition. It is expected that our proposed learning framework would be well extensible for other types of feature fusion, e.g., iris, ear or gait.

Acknowledgment. The work described in this paper was supported by the National Science Foundation of China (No. 61673185, 61502183, 61572205, 61673186), National Science Foundation of Fujian Province (2017J01112), Promotion Program for Young and Middle-aged Teacher in Science and Technology Research (No. ZQN-PY309), the Promotion Program for graduate student in Scientific research and innovation ability of Huaqiao University (No. 1611314014).

References

1. Bredin, H., Chollet, G.: Audio-visual speech synchrony measure for talking-face identity verification. In: Processing of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 233–236 (2007)
2. Cheng, H.T., Chao, Y.H., Yeh, S.L., Chen, C.S.: An efficient approach to multimodal person identity verification by fusing face and voice information. In: Processing of IEEE International Conference on Multimedia and Expo, pp. 542–545, 2005
3. Feng, W., Xie, L., Zeng, J., Liu, Z.Q.: Audio-visual human recognition using semi-supervised spectral learning and hidden markov models. J. Vis. Lang. Comput. **20**(3), 188–195 (2009)
4. Geng, J., Liu, X., Cheung, Y.: Audio-visual speaker recognition via multi-modal correlated neural networks. In: IEEE/wic/acm International Conference on Web Intelligence Workshops, pp. 123–128 (2016)
5. Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional LSTM networks for improved phoneme classification and recognition. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3697, pp. 799–804. Springer, Heidelberg (2005). doi:10.1007/11550907_126
6. Haghighat, M., Abdel-Mottaleb, M., Alhalabi, W.: Discriminant correlation analysis: real-time feature level fusion for multimodal biometric recognition. IEEE Trans. Inf. Forensics Secur. **11**(9), 1984–1996 (2016)
7. Hu, Y., Ren, J.S.J., Dai, J., Yuan, C., Xu, L., Wang, W.: Deep multimodal speaker naming. In: Proceedings of Annual ACM International Conference on Multimedia, pp. 1107–1110 (2015)
8. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceeding of IEEE International Conference on Machine Learning, pp. 448–456 (2015)

9. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Processing of IEEE International Conference on Machine Learning Workshop, pp. 1–6 (2013)
10. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of IEEE International Conference on Machine Learning, pp. 689–696 (2011)
11. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Commun. **54**(4), 543–565 (2012)
12. Soltane, M., Doghmane, N., Guersi, N.: Face and speech based multi-modal biometric authentication. Process. IEEE Int. J. Adv. Sci. Technol. **21**(6), 41–56 (2010)
13. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
14. David Sánchez, A.V.: Advanced support vector machines and kernel methods. Neurocomputing **55**(1C2), 5–20 (2003)