Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Large-scale Multi-modal Person Identification in Real Unconstrained Environments

Jiajie Ye, Yisheng Guan*, Junfa Liu, Xinghong Huang and Hong Zhang

Biomimetic and Intelligent Robotics Lab (BIRL)
School of Electro-mechanical Engineering
Guangdong University of Technology
Guangzhou, Guangdong Province, China

Abstract— **Person identification (P-ID) under real unconstrained noisy environments is a huge challenge. In multiple-feature learning with Deep Convolutional Neural Networks (DCNNs) or Machine Learning method for large-scale person identification in the wild, the key is to design an appropriate strategy for decision layer fusion or feature layer fusion which can enhance discriminative power. It is necessary to extract different types of valid features and establish a reasonable framework to fuse different types of information. In traditional methods, different persons are identified based on single modal features to identify, such as face feature, audio feature, and head feature. These traditional methods cannot realize a highly accurate level of person identification in real unconstrained environments. The study aims to propose a fusion module to fuse multi-modal features for person identification in real unconstrained environments.**

Index Terms— **Multi-modal, fusion strategy, person identification.**

I. INTRODUCTION

Most existing algorithms for person identification are based on quite constrained conditions since the partial datasets used to train are based on a constrained environment. Recently, there is an open-source video dataset based on the real unconstrained environment usually referred to as 'the wild' for person identification, iQIYI. Now, videos play a dominated role in the traditional media or new media since they contain multiple types of information, such as images, audio and text message. Consequently, video understanding has been explored for completing multiple tasks, especially P-ID. The task is mainly solved by face recognition, speaker recognition or any other biometric identification methods. The application scope of P-ID is wide and includes the authentication in high-security systems and forensic tests, and searching for persons in large corpora of video data. All such tasks require high multi-features learning performance under 'real world' conditions.

In video analysis, each topic addresses a single modal of information. As deep learning has been developing rapidly in recent years, all these video understanding methods have achieved great success. In the field of face recognition, based on the LFW benchmark [1], ArcFace [11] realized a precision of 99.83%, which had exceeded the human performance. The best results on Megaface [2] also reached 99.39%. In the field of speaker recognition, the Classification Error Rates of SincNet [12] based on the TIMIT dataset [13] and LibriSpeech dataset [14] were only 0.85% and 0.96%, respectively.

Everything seems right until these P-ID methods are applied in real unconstrained videos. Face recognition is sensitive to pose, blur, occlusion, etc. Moreover, in many video frames, faces are invisible, thus largely increasing the difficulty in face recognition. In person re-identification (Re-ID), the problem of changing clothes has not been considered yet. In the field of speaker recognition, one major challenge comes from the fact that the person to be recognized is not always speaking. Generally speaking, every single technique is inadequate to solve all the cases. Intuitively, the combination of all these sub-tasks together can fully utilize the rich contents of videos [10].

Some open source datasets are available to solve the P-ID task (Table I). The available video datasets mainly utilize a single modality of feature, either face, audio, or body.

Based on different modalities of features, we used a fusion module to fuse multi-modal features for person identification in real unconstrained environments. We investigated different architectures and techniques for training deep CNNs with the features directly extracted from raw video files with little pre-processing and analyzed the performances of several state-of-the-art methods of single modal and multi-modal on iQIYI-VID dataset.

II. RELATED WORKS

A. Face Recognition

The task of face recognition can be divided into two sub-tasks, face verification and face identification. Face verification is a 1-to-1 matching problem of verifying whether the

* Corresponding author, email: ysguan@gdut.edu.cn. This work is partially supported by the National Natural Science Foundation of China (Grant No. 51975126), the Frontier and Key Technology Innovation Funds of Guangdong Province (Grant No. 2017B050506008, 2019B090915001) and the Guangdong Yangfan Program for Innovative and Entrepreneurial Teams (Grant No. 2017YT05G026).

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

TABLE I

DATASETS FOR PERSON IDENTIFICATION. THERE ARE MANY DATASET FOR FACE RECOGNITION AND SOME FOR SPEAKER RECOGNITION, BUT MOST OF THEM ONLY FOCUS ON ONE MODALITY TO COMPLETE THE TASK OF P-ID. RECENTLY, SOME PROJECTS HAVE OPENED SOURCE MULTI-MODAL DATASETS.

Dataset	Task	Identities	Format	Clips/Face tracks	Images/Frames
LFW [1]	Face Recog.	5K	Image	-	13K
Megaface [2]	Face Recog.	690K	Image	-	1M
MS-Celeb-1M [3]	Face Recog.	100K	Image	-	10M
YouTube Celebrities [4]	Face Recog.	47	Video	1,910	-
YouTube Faces [5]	Face Recog.	1,595	Video	3,425	620K
Buffy the Vampire Slayer [6]	Face Recog.	Around 19	Video	12K	44K
Big Bang Theory [7]	Face & Speaker Recog.	Around 8	Video	3,759	-
Sherlock [8]	Face & Speaker Recog.	Around 33	Video	6,519	-
VoxCeleb [9]	Speaker Recog.	6,112	Video	150K	-
iQIYI-VID-2019 [10]	Search	10K	Video	200K	4M

two given images belong to the same person. In 2007, the Labeled Faces in the Wild (LFW) dataset established for face verification by Huang et al. [1] has become the most popular benchmark for face verification. Nevertheless, face recognition is a 1-to-k matching problem of recognizing whether a given image belongs to the image dataset containing k images. Based on LFW, many algorithms [15]–[18] realized the recognition rate above 99%, which was better than the human performance [19]. The state-of-art method, ArcFace [11] achieved a face verification accuracy of 99.83% based on LFW. In this study, ArcFace was adopted to recognize faces in videos.

For the purpose of enriching the contents of features, video datasets are a better choice in multi-modal P-ID. There are some video datasets, such as YouTube celebrity recognition dataset [4] that includes videos of only 35 celebrities, and the YouTube Face Database (YFD) [5] that contains 3425 videos of 1595 persons. The biggest one is iQIYI-VID dataset that even contains some videos without visible faces.

B. Speaker Identification

The applications of speech processing technology are primarily classified as: speech recognition and speaker recognition. Speech recognition is to identify the spoken words, while speaker recognition is to identify speaker on the basis of his/her voice characteristics [20]. Speaker recognition is further dissected into two categories, speaker verification and speaker identification. Speaker verification is the process of validating the claim of identity by a speaker and consequently this type of decision is binary, i.e., true or false. In speaker identification, since there is no prior claim of an identity, the system classifies the input tested speech signals into one of the 'N' reference speakers. Speaker identification stated above is labeled as 'closed-set' speaker identification, which is different from 'open-set' identification, as in the case of open-set, the test speech signal may not belong to any of the 'N' reference speakers and N+1 decisions exist, thus leading to an additional result of the test signal not appertaining to any of the N reference speakers [21].

Currently, speaker recognition still faces a dearth of freely available large-scale datasets in the wild. Some datasets, which were originally intended to be applied in speech recognition, such as TIMIT [13] and LibriSpeech [14], have been adopted in speaker recognition experiments. Many of these datasets were collected under controlled conditions and therefore improper for evaluating models under real conditions. To fill the gap, the Speaker in the Wild (SITW) dataset [22] were generated from open multi-media resources. To the best of our knowledge, the largest and freely available speaker recognition datasets are VoxCeleb [9] and VbxCeleb2 [23].

C. Head Feature Based Hairstyle Classification

Despite audio recognition and face recognition can solve most of the problems of person identification, there are still some loopholes to increase the robustness of existing person identification algorithms. One of the limitations of most existing algorithms is the incapability to detect the presence of human beings under fully unconstrained conditions. Especially, if it is required to detect the presence of human beings from the back or over-the-shoulder views, without clear head-and-shoulder profiles, only human hair or accessories are available. Furthermore, hair contains the characteristics of textures and style and is one of the definite characteristics of human beings since it represents different cultures, historical periods, personal characteristic, gender and age. Hair detection in images is useful for face recognition, person identification, gender classification [24], and head detection in surveillance applications [25].

III. PROPOSED METHOD

The proposed method can be divided into two parts: features extraction and features fusion.

A. *Features extraction*

1) Face: With the PyramidBox [26], one of the best face detectors, we could detect faces in the video frame. To improve the performance of the detector, the VGG16 backbone network of the PyramidBox is replaced with resnet50. In this way, the face recognition model ArcFace [11] can be used for feature extraction with VGG16.

According to our observations, satisfactory results cannot be obtained with only face features. Thus, two coefficients (detection score and quality score) are introduced to obtain the weighted average for the combined prediction. Firstly, the detection score is the highest score ranked based on the confidence of face detector. Some video clips may have more than one face with their correspondent bounding boxes. The bounding box with the biggest size because is selected since the features of only one face are required to predict the identification. Secondly, the quality score is the score simply ranked based on the L2- norm results of the faces output by the FC layer in the face recognition model of SphereFace [27]. Ranjan et al. observed that the L2-norm of the features learned with softmax loss was informative for measuring the face quality [28]. In our experiments, the faces, which were regarded as low-quality faces, were mostly blurred faces, side faces, faces with partial error, and even invisible faces.

2) Audio: All audios from the video clips are first converted into single-channel16-bit streams at a sampling rate of 16 kHz for consistency. Spectrograms are then generated in a sliding window fashion under the hamming window width of 25 ms, a step of 10 ms and 512-point Fast Fourier Transform (FFT). Then, inspired by the previous report [9], these spectrograms are used as the input to the CNN model of VGG-M [29], which is trained as a classification model with the Voxceleb2 dataset [23]. The 512D output from the last hidden layer is used as speaker embedding.

3) Head: Head features are a strong support to increase the robustness of the algorithm of P-ID in the wild and contain information from hair texture, hair style and accessories.

The head detector is YOLOv3 [30]. In some video clips, there are multiple persons mapped to multiple bounding boxes for heads, but only one major character is required. In order to acquire the major character, a major head is defined as a video frame, in which the biggest head is detected. After segmenting the head from a video frame, those head pictures will be resized and normalized as the input to be transferred to a head classifier based on the ArcFace model [11] and the previous report [31].

In general, the feature extraction procedure from face, audio and head are shown in Fig. 1.

B. *Feature fusion*

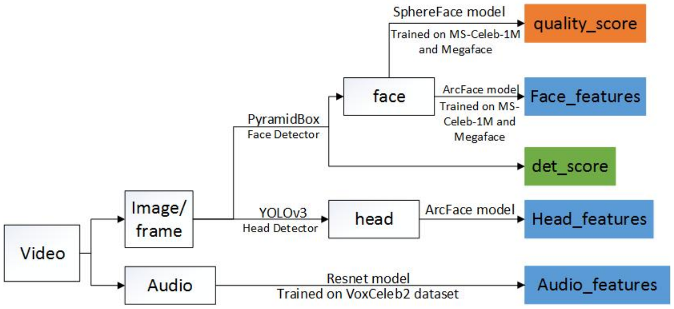


Fig. 1. Feature extraction. It is necessary to extract three representative and effective features with multiple detectors and models trained on different datasets.

The fusion strategy of processing the input videos can be divided into two parts (Part A and Part B). Firstly, a single model of face is utilized to recognize the person identification based on the high-quality score and detection score, as shown in Part A in Fig. 2. Secondly, with the information from face, audio and head, the person identification is recognized in low-quality score and detection score, as shown in Part B of Fig. 2. The details of the fusion strategy are shown in Fig. 2.

- Part A represents the high-score partition and Part B represents the low-score partition.
- According to each video's quality scores and detection scores, the videos with high scores of quality and detection are allocated to Part A for face recognition and other video clips are allocated to Part B.
- In Part A, each video has an unknown number of frames, which contain face features, quality score and detection score. First, through multiplying quality score by detection score, the weight can be obtained, as shown in Eq. (1). Second, all the weights are normalized. Third, with the weight and their corresponding face features, the weighted average can be calculated. Fourth, as shown in Eq. (2), only one feature represents one video by integrating all the frames of one video based on the weighted average and those features are transferred to a classifier of 3-layer Multi-Layer Perception (MLP).

$$Normalize = \begin{pmatrix} a_1 = qua\ score_1 * det\ score_1 \\ a_2 = qua\ score_2 * det\ score_2 \\ \vdots \\ a_n = qua\ score_n * det\ score_n \end{pmatrix} \tag{1}$$

$$F = \frac{\sum_{i=1}^{n} f_i * a_i}{\sum_{i=1}^{n} a_i} \tag{2}$$

where $Normalize$ is normalizing inputs; F is the feature representing a video; n is the number of the video's frame; $a_1, \ldots, a_n$ is the result of multiplying quality score by detection score; $qua\ score$ is the quality score; $det\ score$ is the detection score; f_i is the feature of video frame.

- In Part B, single modal feature is used to recognize the P-ID through 3-layer MLP, and their results are fused to predict the result of low-score videos. The key of recognizing P-ID based on low-score videos is the fusion strategy. After the predicted results of three models are obtained, each predicted result has two parameters: result score and rank score. According to the predicted results, we choose the top 100 predicted results based on the ranked confidence. These 100 predicted results corresponded to 100 videos ID. Consequently, each of the three models has 100 predicted results. As indicated in Eq. (3), according to the label, after multiplying the result score by rank score, the weighted score is acquired. However, a certain video ID is not necessarily contained in predicted results of all the three models at the same time. Thus, the weighted score obtained by fusing the predicted results involves three modals and sometime two modals. This case is similar to the real situation. For example, sometimes we only hear the sound, but we cannot see anyone, indicating the lack of visual features. Finally, as shown in Eq. (4), the final result is arranged according to the weighted score.

$$Label\ i:\ W = \sum_{j=1}^{m} result\ score_j * rank\ score_j \tag{3}$$

$$Result = \begin{pmatrix} label\ 1:\ W_{11}, W_{12}, \ldots, W_{1k} \\ label\ 2:\ W_{21}, W_{22}, \ldots, W_{2k} \\ \vdots \\ label\ N:\ W_{N1}, W_{N2}, \ldots, W_{Nk} \end{pmatrix} \tag{4}$$

where W is the weighted score of a video's ID; $result\ score$ is the confidence to match a label; $rank\ score$ is the sort of result score in top 100; m is the number of the same video IDs in the same label from different modals; N is the number of person ID; k is that the top k results for each person ID.

- After integrating the results of Part A and Part B, the final prediction result of the testing set is obtained.

IV. DATASET

In this study, the iQIYI-VID dataset is selected [10] because it is a large-scale dataset that addresses the problem of multi-modal person identification. Especially, there are video clips without speakers or with only asides. The dataset contains 200 K video clips, which are divided into three parts: 40% for training, 30% for validation, and 30% for testing. The dataset contains about 10, 000 identities. To mimic the real environment of video understanding, distracter videos with unknown person identities which are different from the major identities in the training set are inserted into the validation set and testing set. All the videos are manually labeled and can be used a good benchmark to evaluate person identification algorithms. The video clip duration is in the range of 1 30 seconds with an average of 4.72 second, and the distribution of the number of frames for video clips and the number of videos for labeling are shown in Fig. 4.

Besides, the video clip is a challenge case for face recognition since it includes profile, blur, exposure, occlusion, small face, dark face and invisible face. It is difficult to those faces with only face features. The samples are shown in Fig. 3.

V. EXPERIMENT

A. Experimental Preparation

ArcFace [11] adopted raw features trained on MS-Celeb-1M [3] and Megaface datasets [2], and then input those raw features into a 3 layer Multi-Layer Perception (MLP). The pseudo-code of 3-layer MLP is shown in Table II. Adam optimization algorithm with a learning rate of 0.0008 is adopted.

We combined the training set and validation set together and then divided it into 5 training sets, which were respectively used for training. Finally, 5 models respectively trained on 5 datasets were obtained.

In Part A, the high-score videos are divided into 4 groups according to the intervals of quality score, 40-200, 60-200, 80-200 and 100-200. Hence, the predicted results of Part A is a fusion of 20 models (each of the 5-fold dataset is divided into 4 groups).

In Part B, 3 types of features are used to predict the result under the low-score video frame. Note that, it is not necessary to divide the low-score videos because it is not a valid improvement operation. Hence, the predicted result of Part B is a fusion of 15 models (5 fold dataset for 3 types of feature models).

TABLE II

THE PSEUDO-CODE OF THE 3 LAYER MLP

layer name	input	output	parameter
FC1	512	1024	activation='Relu'
BN1	-	-	batchnormalization,batch size=512
Drop1	-	-	keep-prob=0.5
FC2	1024	1024	activation='Relu'
BN2	-	-	batchnormalization,batch size=512
Drop2	-	-	keep-prob=0.5
FC3	1024	10035	activation='softmax'

B. Evaluation Metrics

Mean Average Precision (MAP) [32] is used to evaluate the retrieval results:

$$MAP(Q) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{m_i} \sum_{j=1}^{n_i} Precision(R_{i,j}) \tag{5}$$

where Q is the number of person IDs; m_i is the number of positive examples for the i-th ID; n_i is the number of positive examples within the top k retrieval results for the i-th ID; $R_{i,j}$ is the set of ranked retrieval results from the

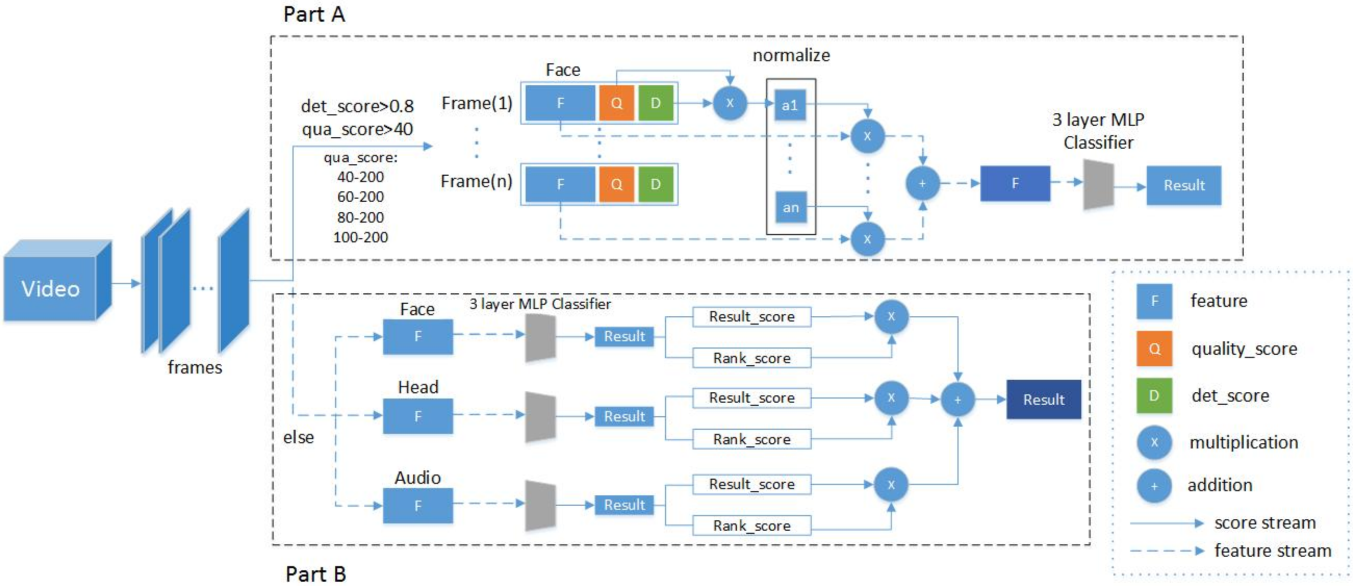


Fig. 2. Fusion strategy. According to the two coefficients of quality score and detection score, the videos with the high synthetic scores are input to the Part A module and the remaining videos are input to the Part B module.

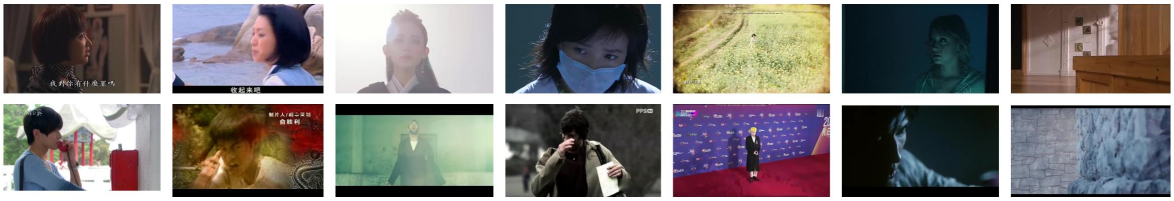


Fig. 3. Challenging cases for face recognition. From left to right: profile, blur, exposure, occlusion, small face, dark face, and invisible face.

top until you get j positive examples. In our implementation, only top 100 retrievals are kept for each person ID.

The top K accuracy is not used in the evaluation since the dataset contains many video clips of unknown identities, which make the top K accuracy invalid.

C. Result

The models were evaluated with the metric of MAP based on the testing set of iQIYI-VID dataset. We compared these models with the state-of-art methods in Table III. Our method achieved a MAP of 92.17%, which was 2.47% higher than that of the current state-of-art method.

Besides, the proposed method integrated multiple models which were trained on the dataset with different quality scores but the same distribution and realized the higher performance than the method, which directly fused 3 types of features and reached a precision of 84.69% in the performance of P-ID.

VI. CONCLUSIONS

In this study, we introduced new architectures and fusion strategies for the task of person identification and demonstrated the state-of-the-art performance based on the iQIYI-VID dataset. Through the comparison with the state-of-art models, experimental results showed that our method fusing multi-modal features outperformed the state-of-art models. Moreover, the provided method, a fusion module to fuse multi-modal features for P-ID in real unconstrained environments, adopts the decision layer fusion based on multiple prediction models, thus improve the accuracy and robustness of P-ID.

REFERENCES

[1] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," *Month*, 2008.

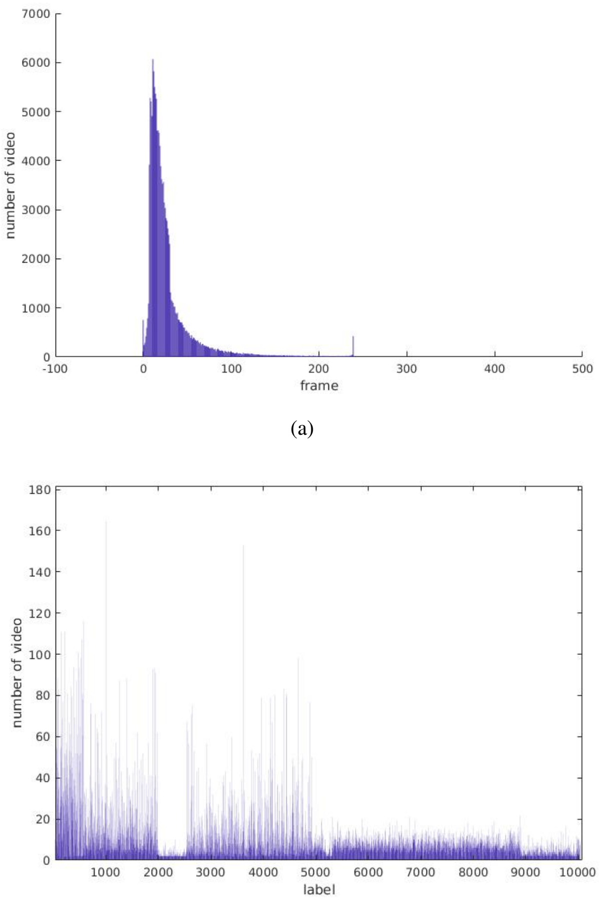
[2] D. Miller, I. Kemelmacher-Shlizerman, and S. M. Seitz, "Megaface: A million faces for recognition at scale," *Computer Science*, 2015.

[3] Y. Guo, Z. Lei, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," 2016.

[4] M. Kim, S. Kumar, V. Pavlovic, and H. A. Rowley, "Face tracking and recognition with visual constraints in real-world videos," 2008.

[5] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision &. Pattern Recognition*, 2011.

[6] J. Sivic, M. Everingham, and A. Zisserman, "who are you? - learning person specific classifiers from video," in *IEEE Conference on Computer Vision &. Pattern Recognition*, 2009.



(a)

(b)

Fig. 4. Distribution of the datasets. (a)The number of frame for a video. (b)The number of video for a label.

[7] M. Bauml, M. Tapaswi, and R. Stiefelhagen, "Semi-supervised learning with constraints for person identification in multimedia data," in *IEEE Conference on Computer Vision &. Pattern Recognition*, 2013.

[8] A. Nagrani and A. Zisserman, "From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script," 2018.

[9] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," 2017.

[10] Y. Liu, P. Shi, B. Peng, H. Yan, Y. Zhou, B. Han, Y. Zheng, C. Lin, J. Jiang, Y. Fan, T. Gao, G. Wang, J. Liu, X. Lu, and D. Xie, "iqiyi-vid: A large dataset for multi-modal person identification," *CoRR*, vol. abs/1811.07548, 2018. [Online]. Available: http://arxiv.org/abs/1811.07548

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," 2018.

[12] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," 2018.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *Nasa Sti/recon Technical Report N*, vol. 93, 1993.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP 2015 - 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," 2015.

TABLE III

COMPARISON WITH STATE-OF-THE-ART METHODS OF SINGLE MODALITY AND MULTI-MODALITIES ON IQIYI DATASET. NOTE THAT, IN THIS EXPERIMENT, MULTIPLE DATASETS WERE USED TO TRAIN THE MODEL AND THIS TASK WAS NOT ACCOMPLISHED JUST BASED ON IQIYI DATASET.

Method		Metric
Modality	Model	MAP(%)
Face	ArcFace-MLP	83.54
Head	ArcFace-MLP	59.42
Audio	Resnet	24.30
Face+Head	Concatenate features and MLP	84.27
Face+Head+Audio	Concatenate features and MLP	84.69
the state-of-art [10]	Attention module	89.70
Ours	Fusion strategies	**92.17**

[16] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *CoRR*, vol. abs/1502.00873, 2015. [Online]. Available: http://arxiv.org/abs/1502.00873

[17] S. Yi, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Computer Vision &. Pattern Recognition*, 2015.

[18] Y. Taigman, Y. Ming, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision &. Pattern Recognition*, 2014.

[19] G. Hu, Y. Yang, Y. Dong, J. Kittler, W. Christmas, S. Z. Li, and T. Hospedales, "When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition," 2015.

[20] J. Campbell, Joseph P., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[21] G. R. Doddington, "Speaker recognitionidentifying people by their voices," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

[22] M. McLaren, L. Ferrer, D. Castn Lavilla, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," 09 2016, pp. 818–822.

[23] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," 2018.

[24] B. Li, X. C. Lian, and B. L. Lu, "Gender classification by combining clothing, hair and facial component classifiers," *Neurocomputing*, vol. 76, no. 1, pp. 18–27, 2012.

[25] Z. Zhang, H. Gunes, and M. Piccardi, "Head detection for video surveillance based on categorical hair and skin colour models," in *IEEE International Conference on Image Processing*, 2009.

[26] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," *CoRR*, vol. abs/1803.07737, 2018. [Online]. Available: http://arxiv.org/abs/1803.07737

[27] W. Liu, Y. Wen, Z. Yu, L. Ming, B. Raj, and S. Le, "Sphereface: Deep hypersphere embedding for face recognition," 2017.

[28] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," 2017.

[29] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *Computer Science*, 2014.

[30] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.

[31] M. Svanera, U. R. Muhammad, R. Leonardi, and S. Benini, "Figaro, hair detection and segmentation in the wild," in *IEEE International Conference on Image Processing*, 2016.

[32] P. R. C. D. Manning and H. Schutze, "Introduction to information retrievalevaluation in information retrieval," 2008.