iQIYI-VID: A Large Dataset for Multi-modal Person Identification

Yuanliu Liu, Peipei Shi, Bo Peng, He Yan, Yong Zhou, Bing Han, Yi Zheng, Chao Lin,
Jianbin Jiang,Yin Fan, Tingwei Gao, Ganwen Wang, Jian Liu, Xiangju Lu, Danming Xie
iQIYI, Inc.

Abstract

Person identification in the wild is very challenging due to great variation in poses, face quality, clothes, makeup and so on. Traditional research, such as face recognition, person re-identification, and speaker recognition, often focuses on a single modal of information, which is inadequate to handle all the situations in practice. Multi-modal person identification is a more promising way that we can jointly utilize face, head, body, audio features, and so on. In this paper, we introduce iQIYI-VID, the largest video dataset for multi-modal person identification. It is composed of 600K video clips of 5,000 celebrities. These video clips are extracted from 400K hours of online videos of various types, ranging from movies, variety shows, TV series, to news broadcasting. All video clips pass through a careful human annotation process, and the error rate of labels is lower than 0.2%. We evaluated the state-of-art models of face recognition, person re-identification, and speaker recognition on the iQIYI-VID dataset. Experimental results show that these models are still far from being perfect for task of person identification in the wild. We further demonstrate that a simple fusion of multi-modal features can improve person identification considerably. We have released the dataset online to promote multi-modal person identification research.

1. Introduction

Person identification is an important task in computer vision. As deep learning arises, face recognition in still images has achieved great success on many benchmarks such as LFW [19] and Megaface [40]. While in real scene videos, the person identification task has not been well explored. First, it is very difficult to collect the video samples, and existing datasets are too small. Second, a single algorithm is not sufficient for solving this problem, we must take full advantage of multi-model information as much as we can to get a satisfactory result. In this paper, we introduce the iQIYI-VID Celebrity Identification dataset, the largest multi-model video person identification dataset, composed of 600K video clips, 5000 celebrity ids, each clip is labeled with only one identity id. Analyzing a large amount of video resources in IQIYI, we utilize multi-modal information in the video to get the person identity, such as face, head, clothes, voice, etc. The whole dataset is manually checked afterwards. As the person identification is the basis of video understanding, we hope this dataset can advance the video person identification research and further accelerate the development of the video understanding task.

Nowadays, videos have dominated the flow on internet. Compared to still images, videos enrich the content by supplying audio and temporal information. As a result, video understanding is of urgent demand for practical usage, and person identification in videos is one of the most important tasks. Person identification has been widely studied in different research areas, including face recognition, person re-identification, and speaker recognition. In the context of video analysis, each research topic addresses a single modal of information. As deep learning arises in recent years, all these techniques have achieved great success. In the area of face recognition, ArcFace [9] reached a precision of 99.83% on the LFW benchmark [19], which had surpass the human performance. The best results on Megaface [40] has also reached 99.39%. For person re-identification (Re-ID), Wang *et al*. [68] raised the Rank-1 accuracy of Re-ID to be 97.1% on the Market-1501 benchmark [82]. In the field of speaker recognition, the Classification Error Rates of SincNet [45] on the TIMIT dataset [49] and LibriSpeech dataset [43] are merely 0.85% and 0.96%, respectively.

Everything seems alright until we try to apply these person identification methods to the real unconstrained videos. Face recognition is sensitive to pose, blur, occlusion, *etc*. Moreover, in many video frames, the faces are even invisible. Re-ID has not touched the problem of changing clothes yet. For speaker recognition, one major challenge comes from the fact that the person to recognize is not always speaking. Generally speaking, every single technique is inadequate to cover all the cases. Intuitively, it will be beneficial to combine all these sub-tasks together, so we can fully utilize the rich content of videos.

There are several datasets for person identification in the

Table 1: Datasets for person identification.

Dataset	Task	Identities	Format	Samples
LFW [19]	Recognition	5K	Image	13K
Megaface [40]	Recognition	690,572	Image	1M
MS-Celeb-1M [12]	Recognition	100K	Image	10M
YouTube Celebrities [25]	Recognition	47	Video	1,910
YouTube Faces [74]	Recognition	1,595	Video	3,425
Market1501 [82]	Re-ID	1,501	Image	32,668
Cuhk03 [28]	Re-ID	1,467	Image	13,164
iLIDS [71]	Re-ID	300	Video	600
Mars [81]	Re-ID	1,261	Video	20K
CSM [21]	Search	1,218	Video	127K
iQIYI-VID	Search	5,000	Video	600K

literature. We list the popular datasets in Table 1. Most of the video datasets focus on only one modal of feature, either face [25, 74] or full body [71, 81, 21]. To our best knowledge there is no large-scale dataset that addresses the problem of multi-modal person identification.

Here we present the iQIYI-VID dataset, which is the first video dataset for multi-model person identification. It contains 600K video clips of 5,000 celebrities, which is the largest celebrity identification dataset. All the videos are manually labeled, which makes it a good benchmark to evaluate person identification algorithms. This dataset aims to encourage the development of multi-modal person identification.

In Section 2, we will review related works of person identification briefly. We present the iQIYI-VID dataset in Section 3. Our baseline method will be given in Section 4. In Section 5, we evaluate our baseline and related methods on our dataset. We conclude our work in Section 6.

2. Related Work

2.1. Face Recognition

Typically, face recognition consists of two tasks, face verification and face identification. Face verification is to determine whether two given face images belong to the same person. In 2007, the Labeled Faces in the Wild (LFW) dataset was built for face verification by Huang *et al*. [19]. The LFW database includes 13K photos of 5K different people, and it is the most widely used benchmark for verification. After that, a large number of algorithms were proposed to solve the problem of face verification, and many of them [50, 56, 58, 59, 60] have reached recognition rates of over 99.9% on LFW, which is better than human performance [18].

In recent years, the interest in face identification has greatly increased. The task of face identification is to find the most similar faces between gallery set and query set. Each individual in gallery set only has some typical face images (less than five). Currently, the most famous and difficult benchmarks are Megaface [40] and MS-Celeb-1M database [12]. Megaface includes one million unconstrained and multi-scaled photos of ordinary people collected from Flickr [2]. The MS-Celeb-1M database provides one million face images of celebrities selected from freebase with corresponding entity keys. Both datasets are large enough for training. However, contents in these two datasets are still images. They are not suitable for training models that recognize faces in unconstrained videos. Moreover, these datasets are noisy as pointed out by [66]. In 2008, Kim *et al*. created YouTube celebrity recognition database [25] that includes only 35 celebrities, mostly actors/actresses and politicians, from YouTube. The Youtube Face Database(YFD) [74] contains 3425 videos of 1595 different people. Both databases have much less videos than iQIYI-VID. Another difference is that our database includes video clips without visible faces.

With the successful application of deep learning in computer vision, the algorithms based on convolutional neural network (CNN) have gradually become the mainstream of face recognition. Since 2014, Yi Sun *et al*. proposed DeepID series [56, 58, 57, 55] based on contrastive loss. In particular, DeepID3 [56] added joint face identification-verification supervision to both intermediate and final feature extraction layers during training, which achieved an accuracy of 99.53% on the LFW dataset. FaceNet [50] adopted the triplet loss. It generated triples that directly learnt a mapping from face images to a compact Euclidean space, where distances directly corresponded to a measure of face similarity. An alternative is the center loss [73], which makes each class more compact in the feature space. The latest face recognition algorithms [33, 32, 34, 67, 9] focused on improving angular/cosine margin-based loss, which made learned features potentially separable with a larger angular/cosine distance. The state-of-art method, ArcFace [9] achieved a face verification accuracy of 99.83%

on LFW. In this paper we take ArcFace to recognize faces in videos.

Face detection is a fundamental step for face recognition. Thanks to the development of deep learning, face detection has achieved great successes in recent years. Recent CNN-based face detectors can be categorized into single-stage detectors [20, 76, 13, 35, 78, 70, 42, 61] and two-stage detectors [69, 83, 29, 72]. There are also cascaded detectors [27, 77] that achieve a very fast speed. We choose SSH [42] as the basic detector, which is the best open-source face detector at this time.

2.2. Audio-based Speaker Identification

Speaker recognition has been an active research topic for many decades [52]. It comes in two forms that speaker verification [5] aims to verify if two audio samples belong to the same speaker, whereas speaker identification [62] predicts the identity of the speaker given an audio sample. Speaker recognition falls into two categories, text dependent [41] and text independent [42]. Speaker recognition has been approached by applying a variety of machine learning models [3, 4, 10], either standard or specifically designed, to speech features such as MFCC [38]. For many years, the GMM-UBM framework [47] dominated this particular field. Recently i-vector [8] and some related methodologies [24, 7, 23] emerged and became increasingly popular. More recently, d-vector based on deep learning [26, 15, 63] have achieved competitive results against earlier approaches on some datasets. Nevertheless, speaker recognition still remains a challenge, especially for data collected in uncontrolled environment and from heterogeneous sources.

Currently, there are not many freely available datasets for speaker recognition, especially for large-scale ones. National Institute of Standards in Technology (NIST) have hosted several speaker recognition evaluations. However, the associated datasets are not freely available. There are datasets originally intended for speech recognition, such as TIMIT [49] and LibriSpeech [43], which have been used for speaker recognition experiments. Many of these datasets were collected under controlled conditions and therefore are improper for evaluating models in real-world conditions. To fill the gap, the Speaker in the Wild (SITW) dataset [39] were created from open multi-media resources and freely available to the research community. To the best of our knowledge, the largest, freely available speaker recognition datasets are VoxCeleb [41] and VoxCeleb2 [6]. They were collected using a completely automatic pipeline and therefore could scale to thousands of speakers. More importantly, similar to SITW datset, the VoxCeleb datasets were created from YouTube videos, ensuring their data diversity in terms of data quality and background noise.

2.3. Body Feature Based Person Re-Identification

Person re-identification (Re-ID) recognizes people across cameras, which is suitable for videos that have multiple shots switching between cameras. In recent years, Re-ID has made a lot of breakthroughs. For single-frame-based methods, the model mainly extracts features from a still image, and directly determines whether the two pictures belong to the same person [11, 30]. In order to improve the generalization ability, character attributes of the person was added to the network [30]. Metric learning measures the degree of similarity by calculating the distance between pictures, focusing on the design of the loss function [64, 51, 16]. These methods rely on the global feature matching on the whole image, which is sensitive to the background. To solve this problem, local features gradually arises. Various strategies are proposed to extract local features, including image dicing, skeleton point positioning, and attitude calibration [65, 80, 79]. Image dicing has the problem of data alignment [65]. Therefore, some researchers proposed to extract the pose information of the human body through skeleton point detection and introduce STN for correction [80]. Zhang *et al*. [79] further calculated the shortest path between local features and introducing a mutual learning approach, so that the accuracy of recognition exceeds that of humans for the first time.

The information within a single frame is often limited. In comparison, video sequence-based methods can extract temporal information to aid Re-ID. AMOC [31] adopted two sub-networks to extract the features of global content and motion. Song *et al*. [54] adopted self-learning to find out low-quality image frames and reduce their importance, which made the features more representative.

Existing video-based Re-ID datasets, such as DukeMTMC [48], iLIDS [71] and MARS [81], have invariant scenes and small amounts of data. The full body are usually visible, and the clothes are always unchanged for the same person crossing cameras. In comparison, video clips of iQIYI-VID are extracted from the massive video database of iQIYI, which covers various types of scenes and the same id spans multiple kinds of videos. Significant changes in character appearances make Re-ID more challenging but closer to practical usage. The image resolution is also much higher than the other datasets.

3. iQIYI-VID DATASET

The iQIYI-VID dataset is a large-scale benchmark for multi-modal person identification. To make the benchmark close to real applications, we extract video clips from real online videos of extensive types. Then we label all the clips by human annotators. Automatic algorithms are adopted to accelerate the collection and labeling process. The flow chart of the construction process is shown in Figure 1. We

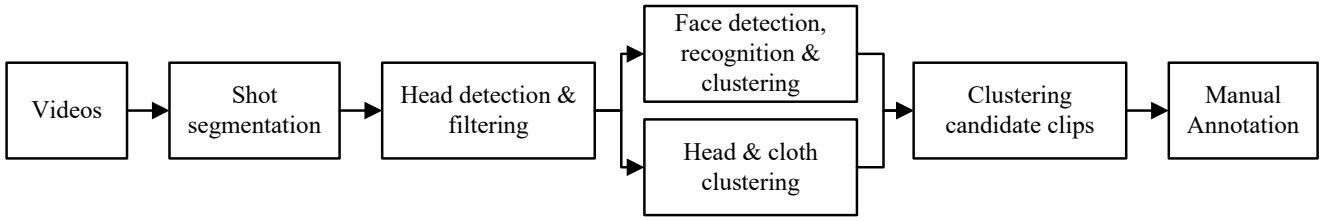


Figure 1: The process of building the iQIYI-VID dataset.

begin with extracting video clips from a huge database of long videos. Then we filter out those clips with no person or multiple persons by automatic algorithms. After that we group the candidate clips by identity and put them into manual annoation. The details are given below.

3.1. Extracting video clips

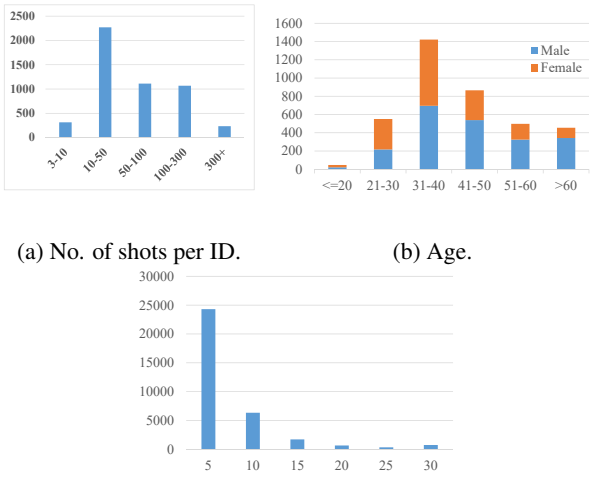
The raw videos are the top 500,000 popular videos of iQIYI, covering movies, teleplays, variety shows, news broadcasting, and so on. Each raw video is segmented into shots according to the dissimilarity between consecutive frames. Video clips that are shorter than one second are excluded from the dataset, since they often contain not enough multi-modal information. Those clips longer than 30 seconds are also dropped out due to a large computation burden.

3.2. Automatic filtering by head detection

As a benchmark for person identification, each video clip is required to contain one and only one major character. In order to find out the major characters, heads are detected by YOLO V2 [46]. A valid frame is defined as a frame in which only one head is detected, or the biggest head is three times larger than the other heads. A valid clip is defined as a clip whose valid frames exceed a ratio of 30%. Invalid clips are dropped out. Since the head detector cannot detect all the heads, some clips will survive in this stage. Such noise clips will be thrown away in the manual filtering step.

3.3. Obtaining candidate clips for each identity

In this stage, each clip will be labeled with an initial identity by face recognition or clothes clustering. All the identities are selected from the celebrity database of iQIYI. The faces are detected by the SSH model [42], and then recognized by the ArcFace model [9]. After that, the clothes and head information are used to cover those clips that no face has been detected or recognized. We cluster the clips from the same video by the faces and clothes information. The face and clothes in each frame are paired according to their relative position, and the identities of faces are propagated to the clothes with the face-clothes pairs. After that, each clothes cluster can get an ID through majority voting. The clips falling into the same cluster are regarded to be



(a) No. of shots per ID. (b) Age.

(c) Shot length (in seconds).

Figure 2: Data distributions.

with the same identity, so the clips with no recognizable face can inherit the label from their clusters. Note that, the bounding boxes of faces and clothes are not included in the published dataset, which matches the case in real applications.

3.4. Final manual filtering

At this point, we have obtained all the candidate clip clusters with noise. All clusters of clips are cleaned by a manual annotation process. More labeling efforts are put on the clips with low scores in face recognition and those been extracted by clothes clustering.

The manual labeling was repeated twice by different labelers to ensure a high quality. After data cleaning, we randomly selected 10% of the dataset for quality testing and the data labeling error rate was kept within 0.2%.

4. Statistics

IQIYI-VID dataset contains 600K videos, which is much larger than the datasets listed in Table 1. The whole dataset is divided into three parts, 40% for training, 30% for validation, and 30% for test. Researchers can download the train-

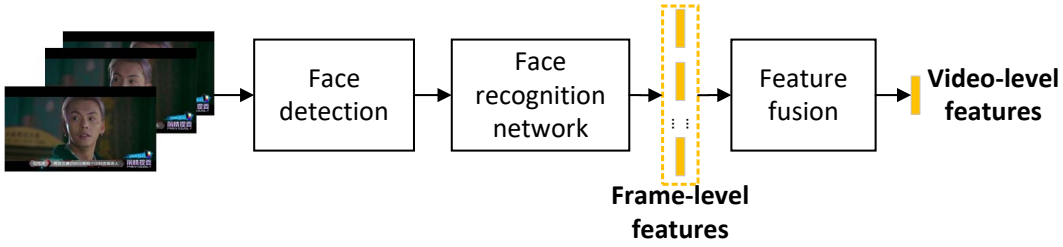


Figure 3: Flowchart for face-based video representation.

ing and validation parts from the website [1], and the test part is kept secret for evaluating the performance of contesters. The video clips are labeled with 5,000 identities from the iQIYI celebrity database. Among all the celebrities, 4,360 of them are Asian, 510 are Caucasian, 41 are African, and 23 are Hispanic. To mimic the real environment of video understanding, we add 84,759 distracter videos with unknown person identities (outside the 5,000 major identities in training set) into the validation set and another 84,746 distracter videos into the test set.

The number of video clips for each celebrity varies between 20 and 900, with 100 on average. The histogram of the number of video clips over ID is shown in Figure 2a. As shown in Figure 2c, the video clip duration is in the range of 1 30 seconds, with 4.72 seconds on average.

The dataset is approximately gender-balanced, with 54% males and 46% females. The age range of the dataset is quite considerable, the earliest birth date was 1877 and the latest birth date was 2012. The histogram is shown in Figure 2b.

5. Baseline Approaches

In this section we present a baseline approach for multi-modal person identification on the iQIYI-VID dataset. Our video processing flow takes a video as input, extracts multi-modal features on single frames separately, and fuses them into video-level features. Then we train classifiers on these video-level features. Video retrieval is finally realized by ranking test videos according to the probability distribution over the IDs.

5.1. Face recognition

As shown in Figure 3, the whole procedure for face recognition includes three main modules, including face detection, face recognition, and feature fusion.

Face Detection. We use the SSH model [42] to detect faces, which is the best open-source face detector. To accelerate the detection, we replace the VGG16 backbone network [53] of the original SSH with MobileNetV1 [17].

[1] http://challenge.ai.iqiyi.com/detail?raceId=5afc36639689443e8f815f9e

Face Recognition and Quality Assessment. We choose the state-of-art face recognition model ArcFace [9]. We also found that the face quality is crucial for face verification and identification performance in practice. So we train a face quality classification model. The classification model can help us make full use of good faces and suppress the interference of bad faces. In our experiment, the face quality is categorized into four types:

1. Blurred face: The image of face is blurred. In many cases these face images cannot be detected nor recognized reliably.

2. Good face: Faces are clear, and face angles are less than 60 degrees.

3. Medium face: It contains not only acceptable face images, but also includes side faces and partial misdetected faces.

4. Side face: Face angles are greater than 60 degrees.

Feature Fusion. In the literature there are many aggregation techniques for getting video-level representations from frames, such as Fisher Vectors (FV) [44] and VLAD [22]. But those techniques are task-independent, which do not consider face quality in this case. In contrast, we take the cluster center of face features as the video-level presentations based on face quality levels. We sort face quality in the order of good face, side face, medium face, and blurred face. Firstly, if a video contains good face, we take the cluster center of good faces as the video-level features. Then if a video does not contain any good face but has side faces, we only take side face into account. The medium faces and blurred faces are examined in turn.

5.2. Multi-modal features

We utilize the state-of-art models to extract multi-modal features, including head, audio, and body. The procedure is similar to face recognition in Section 5.1.

Head Recognition. We train a head classifier based on the ArcFace model [9]. The heads are detected by YOLO V2 [46] as mentioned in Section 3.2. The head features

contain information from hair style and accessories, which are good supplement to faces.

Audio. The audio from the video clips is converted to a single-channel, 16-bit signal with a 16kHz sampling rate. The magnitude spectrograms of these audio are mean-subtracted, but variance normalisation is not performed. Instead, we only divide those mean-subtracted spectgrograms by a constant value 50.0 and feed the results as input to a CNN model based on ResNet34 [14], inspired by [40]. The CNN model is trained as a classification model using the dev part of Voxceleb2 dataset [4] with 5994 speakers, 14% of the data is used as evaluation while the rest as training data. We achieved a best classification error rate of 6.3% on the evaluation data. The 512D output from the last hidden layer is used as speaker embedding.

Body. We utilize Alignedreid++ [36] to recognize the person by its body feature. We detect the persons by a SSH detector [42] trained on the COCO dataset [1].

5.3. Model fusion

To utilize multi-modal information for person recognition, we adopt an early-fusion strategy to combine the features extracted in Section 5.1 and 5.2. We extract features from the single models and fuse them simply by a weighted sum. In our implementation, the feature layer right before the softmax layer is fed to the fusion. We set the feature lengths of all the models to be exactly the same, so they can be added up.

6. Experiments

6.1. Experiment setup

The modified SSH face detector is trained on the Widerface dataset [75]. The face quality classification model in Section 5.1 is trained on our private dataset. In Section 5.3, the feature length is set to 512. The weights for face, head, audio, and body are empirically set to 1, 0.5, 0.5, and 0.2, respectively. The underlying strategy is that more reliable features gains higher weights.

6.2. Evaluation Metrics

To evaluate the retrieval results, we use Mean Average Precision (MAP) [37]:

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

where Q is the set of person IDs to retrieve, n_i is the number of positive examples within the top k retrieval results for the i-th ID, and $R_{i,j}$ is the set of ranked retrieval results from the top until you get j positive examples. In our implementation, only top 100 retrievals are kept for each person ID.

Table 2: Results of different modal of features.

Modal	MAP (%)
Face	85.19
Head	54.32
Audio	11.79
Body	5.14
Face+Head	87.16
Face+Head+Audio	87.69
Face+Head+Audio+Body	87.80
Combine 3 models	89.24
Combine 6 models	89.35

6.3. Results on iQIYI-VID

Figure 4: Challenging cases for head recognition. The hair style and accessories of the same actor changed dramatically.

(a) (b)

Figure 5: Challenging cases for body recognition. (a) An actress changes style in different episodes. (b) Different actors dress the same uniform.

We evaluate the models described in Section 5 on the test set. The results are given in Table 2. Among all the models using single modal of feature, face recognition achieved the best performance. It should be mentioned that, ArcFace [9] achieved a precision of 99.83% on the LFW benchmark [19], which is much higher than on the iQIYI-VID dataset. It suggested that, the iQIYI-VID dataset is more challenging than LFW. In Figure 7 we show some difficult examples that are hard to recognize using only face features.

None of the other features alone is comparable to face features. The head feature is unreliable when the face is hidden. In this case the classifier mainly rely on hair or accessories, which are much less discriminative than face fea-

Figure 6: An example of video retrieval results. When adding more and more features inside, the results get much better than face recognition alone. The number above each image indicates the rank of the image within the retrieval results. Positive examples are marked by green boxes, while negative examples are marked in red.

Figure 7: Challenging cases for face recognition. From left to right: profile, occlusion, blur, unusual illumination, and small face.

tures. Moreover, the actors or actresses often change their hairstyles across the shows, as shown in Figure 4.

The MAP of audio-based model is only 11.79%. The main reason is that the person identity of the video clip is primarily determined by the figure in the video frame, we didn't use a voice detection module to filter out non-speaking clips, nor was active speaker detection employed. As a result, the sound may likely not come from the person of interest. Moreover, in many cases, even though the person of interest does speak, the voice actually comes from a voice actor for that person, which makes recognition by speaker more problematic. When the character does not say anything inside the clip and the audio may come from the background, or even from some other characters, which are distracters for speaker recognition.

The performance of body feature is even worse. The main challenge comes from two aspects. In the one hand, the clothes of the characters always change from one show to another. In the other hand, the uniforms of different characters may be nearly the same in some shows. Figure 5 gives an example. As a result, the intra-class variation is comparable to, if not larger than, the inter-class variation.

Although neither head, audio, nor body feature alone can recognize person well enough, they can produce better classifiers when combined with the face feature. From Table 2 we can see that, adding more features always achieves bet-

ter performances. Adopting all four kinds of features can raise the MAP by over 2.61%, from 85.19% to 87.80%. It proves that multi-modal information is important for video-based person identification. An example is shown in Figure 6. We can see that when multi-modal features are added to person recognition gradually, more and more positive examples are retrieved back.

Inspired by the strategy of Cross Validation, we built a simple expert system that is composed of models trained on different partition of the training set. The output of the expert system is the sum of the output probability distribution of all the models. When using three models, the MAP can be raised by 1.44% further, from 87.8% to 89.24%. However, using more experts will not increase the performance much. As shown in Table 2, using six models, the MAP can be raised only 0.11%, that is, from 89.24% to 89.35%.

7. CONCLUSIONS AND FUTURE WORK

In this work, we investigated the challenges of person identification in real videos. We built a large-scale video dataset called iQIYI-VID, which contains more than 600K video clips and 5,000 celebrities extracted from iQIYI copyrighted videos. Our baseline approach of multi-modal person identification demonstrated that it is beneficial to take different sources of features to deal with real-world videos. We hope this new benchmark can promote the research in multi-modal person identification.

References

[1] Coco dataset. `http://cocodataset.org/`.

[2] Flickr. `https://www.flickr.com/`.

[3] S. G. Bagul and R. K. Shastri. Text independent speaker recognition system using gmm. In *International Conference on Human Computer Interactions*, pages 1–5, 2013.

[4] A. Bajpai and V. Pathangay. Text and language-independent speaker recognition using suprasegmental features and support vector machines. In *International Conference on Contemporary Computing*, pages 307–317, 2009.

[5] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, and D. A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP J. Adv. Sig. Proc.*, 2004(4):430–451, 2004.

[6] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, pages 1086–1090, 2018.

[7] S. Cumani, O. Plchot, and P. Laface. Probabilistic linear discriminant analysis of i-vector posterior distributions. In *International Conference on Acoustics, Speech and Signal Processing*, pages 7644–7648, 2013.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Trans. Audio, Speech & Language Processing*, 19(4):788–798, 2011.

[9] J. Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CoRR*, abs/1801.07698, 2018.

[10] S. W. Foo and E. G. Lim. Speaker recognition using adaptively boosted classifier. In *International Conference on Electrical and Electronic Technology*, 2001.

[11] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *CoRR*, abs/1611.05244, 2016.

[12] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102, 2016.

[13] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu. Scale-aware face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1913–1922, 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer. End-to-end text-dependent speaker verification. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5115–5119, 2016.

[16] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[18] G. Hu, Y. Yang, D. Yi, J. Kittler, W. J. Christmas, S. Z. Li, and T. M. Hospedales. When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition. In *IEEE International Conference on Computer Vision Workshop*, pages 384–392, 2015.

[19] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[20] L. Huang, Y. Yang, Y. Deng, and Y. Yu. Densebox: Unifying landmark localization with end to end object detection. *CoRR*, abs/1509.04874, 2015.

[21] Q. Huang, W. Liu, and D. Lin. Person search in videos with one portrait through visual and temporal links. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, pages 437–454, 2018.

[22] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1704–1716, 2012.

[23] P. Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. Technical report, 2005.

[24] P. Kenny. Bayesian speaker verification with heavy-tailed priors. In *Odyssey 2010: The Speaker and Language Recognition Workshop*, page 14, 2010.

[25] M. Kim, S. Kumar, V. Pavlovic, and H. A. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[26] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu. Deep speaker: an end-to-end neural speaker embedding system. *CoRR*, abs/1705.02304, 2017.

[27] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5325–5334, 2015.

[28] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition pages = 152–159, year = 2014*.

[29] Y. Li, B. Sun, T. Wu, and Y. Wang. Face detection with end-to-end integration of a convnet and a 3d model. In *European Conference on Computer Vision*, pages 420–436, 2016.

[30] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *CoRR*, abs/1703.07220, 2017.

[31] H. Liu, Z. Jie, J. Karlekar, M. Qi, J. Jiang, S. Yan, and J. Feng. Video-based person re-identification with accumulative motion context. *CoRR*, abs/1701.00193, 2017.

[32] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6738–6746, 2017.

[33] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*, pages 507–516, 2016.

[34] W. Liu, Y. Zhang, X. Li, Z. Liu, B. Dai, T. Zhao, and L. Song. Deep hyperspherical learning. In *Advances in Neural Information Processing Systems*, pages 3953–3963, 2017.

[35] Y. Liu, H. Li, J. Yan, F. Wei, X. Wang, and X. Tang. Recurrent scale approximation for object detection in CNN. In *IEEE International Conference on Computer Vision*, pages 571–579, 2017.

[36] H. Luo, W. Jiang, X. Zhang, X. Fan, Q. Jingjing, and C. Zhang. Alignedreid++: Dynamically matching local information for person re-identification. 2018.

[37] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[38] J. Martínez, H. Pérez-Meana, E. E. Hernández, and M. M. Suzuki. Speaker recognition using mel frequency cepstral coefficients (MFCC) and vector quantization (VQ) techniques. In *International Conference on Electrical Communications and Computers*, pages 248–251, 2012.

[39] M. McLaren, L. Ferrer, D. Castán, and A. Lawson. The speakers in the wild (SITW) speaker recognition database. In *Interspeech*, pages 818–822, 2016.

[40] D. Miller, I. Kemelmacher-Shlizerman, and S. M. Seitz. Megaface: A million faces for recognition at scale. *CoRR*, abs/1505.02108, 2015.

[41] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Interspeech*, pages 2616–2620, 2017.

[42] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis. SSH: single stage headless face detector. In *IEEE International Conference on Computer Vision*, pages 4885–4894, 2017.

[43] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210, 2015.

[44] F. Perronnin and C. R. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[45] M. Ravanelli and Y. Bengio. Speaker recognition from raw waveform with sincnet. *CoRR*, abs/1808.00158, 2018.

[46] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6517–6525, 2017.

[47] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[48] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*, pages 17–35, 2016.

[49] J. S. Garofolo, L. Lamel, W. M. Fisher, J. Fiscus, and D. S. Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. 93:27403, 01 1993.

[50] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. pages 815–823, 03 2015.

[51] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[52] C. D. Shaver and J. M. Acken. A brief review of speaker recognition technology. *Electrical and Computer Engineering Faculty Publications and Presentations*, 2016.

[53] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[54] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai. Region-based quality estimation network for large-scale person re-identification. In *AAAI*, 2018.

[55] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014.

[56] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *CoRR*, abs/1502.00873, 2015.

[57] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10, 000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[58] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2892–2900, 2015.

[59] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[60] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2746–2754, 2015.

[61] X. Tang, D. K. Du, Z. He, and J. Liu. Pyramidbox: A context-assisted single shot face detector. In *European Conference on Computer Vision*, pages 812–828, 2018.

[62] R. Togneri and D. Pullella. An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits and Systems Magazine*, 11:23–61, 2011.

[63] E. Variani, X. Lei, E. McDermott, I. Lopez-Moreno, and J. Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *International Conference on Acoustics, Speech and Signal Processing*, pages 4052–4056, 2014.

[64] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808, 2016.

[65] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153, 2016.

[66] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. C. Loy. The devil of face recognition is in the noise. In *European Conference on Computer Vision*, pages 780–795, 2018.

[67] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Process. Lett.*, 25(7):926–930, 2018.

[68] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularity for person re-identification. In *ACM Multimedia*, 2018.

[69] H. Wang, Z. Li, X. Ji, and Y. Wang. Face R-CNN. *CoRR*, abs/1706.01061, 2017.

[70] J. Wang, Y. Yuan, and G. Yu. Face attention network: An effective face detector for the occluded faces. *CoRR*, abs/1711.07246, 2017.

[71] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703, 2014.

[72] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li. Detecting faces using region-based fully convolutional networks. *CoRR*, abs/1709.05256, 2017.

[73] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016.

[74] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–534, 2011.

[75] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[76] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. S. Huang. Unitbox: An advanced object detection network. In *ACM Multimedia*, pages 516–520, 2016.

[77] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016.

[78] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S^3fd: Single shot scale-invariant face detector. In *IEEE International Conference on Computer Vision*, pages 192–201, 2017.

[79] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. Alignedreid: Surpassing human-level performance in person re-identification. *CoRR*, abs/1711.08184, 2017.

[80] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 907–915, 2017.

[81] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. MARS: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, 2016.

[82] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *International Conference on Computer Vision*, pages 1116–1124, 2015.

[83] C. Zhu, Y. Zheng, K. Luu, and M. Savvides. CMS-RCNN: contextual multi-scale region-based CNN for unconstrained face detection. *CoRR*, abs/1606.05413, 2016.