Multi-Cue and Temporal Attention for Person Recognition in Videos

Wenzhe Wang, Bin Wu(✉), Fangtao Li, and Zihe Liu

Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China
{wangwenzhe,wubin,lift,ziheliu}@bupt.edu.cn

Abstract. Recognizing persons under unconstrained settings is challenging due to variation in pose and viewpoint, partial occlusion, and motion blur. Inference only by face-based recognition techniques would fail in these cases. Previous studies mainly focus on this problem on still images while they cannot handle the temporal variations in videos. In this work, we aim to tackle these challenges and propose a Multi-Cue and Temporal Attention (MCTA) framework to recognize persons in videos. For the spatial domain, we extract features from multiple visual cue regions and utilize a Multi-Cue Attention Module to integrate them. For the temporal domain, we adopt a Temporal Attention Module to combine the video frames, which is learned to assess the quality of different frames adaptively. By this means, MCTA can comprehensively explore the complementary information in spatial-temporal dimensions for person recognition in videos. Moreover, we introduce Character Recognition in Videos (CRV), a new video dataset for character recognition under challenging settings. Extensive experiments on CRV demonstrate the effectiveness of our proposed framework. Dataset with annotations and all codes used in this paper are publicly available at https://github.com/zhezheey/MCTA.

Keywords: Person recognition · Multiple cues · Spatial-temporal attention

1 Introduction

Recognizing persons in images or videos is frequently needed in practical scenarios. As a key task of the understanding of multimedia content and computer vision, person recognition has been widely studied and achieved great success in multiple settings, including face recognition [1,2], person re-identification [3,4], and speaker recognition [5]. Nonetheless, person recognition under unconstrained scenarios remains a challenging task and far from being well solved. Issues like great variation in pose and viewpoint, partial occlusion, motion blur, and noisy sounds in practice bring substantial difficulties for these methods.

Supported by the National Key R&D Program of China (2018YFC0831500), the National Natural Science Foundation of China (No. 61972047), and the NSFC-General Technology Basic Research Joint Funds (No. U1936220).

Y. Peng et al. (Eds.): PRCV 2020, LNCS 12306, pp. 369–380, 2020.
https://doi.org/10.1007/978-3-030-60639-8_31

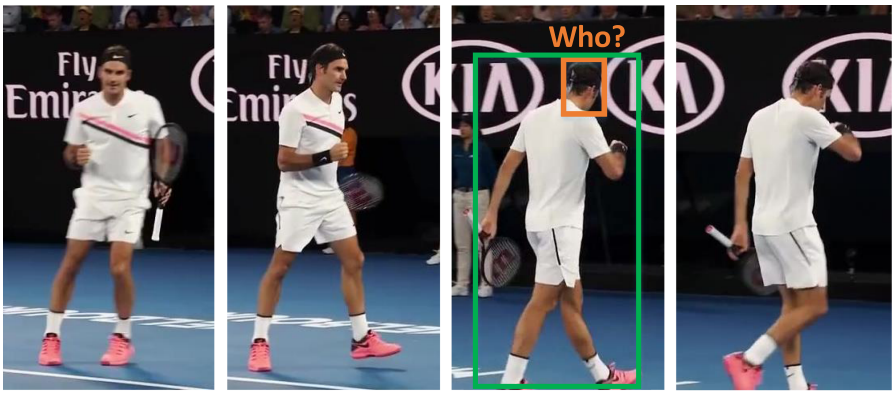

Fig. 1. How do we recognize a person in videos when the face is invisible? The hairstyle, clothing, scene, and information of adjacent frames are significant.

To tackle the problem, a natural idea is to combine the information of multiple cues. Existing studies mainly focus on image-based condition, in which additional visual cues, such as hairstyles, clothing, or scenes, are utilized to recognize persons in photos when the faces are blurred or even cropped. They rely on concatenation [6,7], heuristics [8,9], or simple attention methods [10] to combine contextual cues from multiple regions and achieve better results on benchmarks.

However, compared to image-based person recognition, the video-based scenario attracts far less attention. Recognizing persons in unconstrained videos is a more challenging task with many practical applications, which needs both the spatial and temporal context to help recognition (see Fig. 1). Recently, iQIYI and ACM Multimedia held a challenge [11] towards multi-model person identification based on a large-scale video dataset and attracted hundreds of researchers [12–15] to come up with novel ideas. However, these studies are limited in two aspects: (1) The large dataset with high-quality faces (99.65% of videos contain clear faces) provides much richer information than practical scenarios that the model can get over 90% mAP only by face features [12], making other visual and multi-model cues almost useless. (2) These methods rely on simple concatenation [14,15] to integrate multi-cue information, and averaging [12] or heuristic rules [13,15] to model the temporal information separately, which are obviously over-simplified.

In this work, we propose a Multi-Cue and Temporal Attention (MCTA) framework for person recognition in videos. In the spatial domain, different regions of a person, including face and upper body, are detected and then input along with the whole image into specific Convolutional Neural Networks (CNNs) to extract the features. Moreover, we adopt a Multi-Cue Attention Module (MCAM) to adaptively combine multiple visual cues. In the temporal domain, we introduce a Temporal Attention Model (TAM), which applies an attention weighted average on the sequence of image features, to model the importance of

different frames. Finally, our MCTA framework achieves person recognition by end-to-end spatial-temporal information learning in videos.

To evaluate our method under more diverse settings and facilitate the research, we construct a dataset named Character Recognition in Videos (CRV) by annotating 79 characters in 4,405 video clips from 34 movies or TV series. In particular, all of the actors or actresses play multiple roles in one movie or TV series, and there are certain differences between the characters. To simplify the problem, each video clip is limited to contain only one or one main person similar to [11]. Compared with the existing dataset [11], CRV is more challenging, which requires additional visual cues like hairstyles or clothing to complement facial information and recognize the characters. Experiments demonstrate that our proposed framework can significantly raise the performance on CRV.

In summary, the main contributions of this work include:

- We propose a novel MCTA framework to recognize persons in videos under unconstrained settings, which incorporates both the spatial and temporal information for person recognition.
- We introduce an MCAM to integrate features of visual cues from multiple regions, and a TAM to access the quality of different frames and combine them. These two modules together comprehensively explore the information of persons in videos.
- We construct CRV, a novel and challenging dataset for character recognition, to promote the research on person recognition in videos.
- Our framework achieves state-of-the-art performance on CRV, which demonstrates the effectiveness of our method.

2 Related Work

2.1 Face Recognition and Person Re-identification

As the most widely studied and applied direction of person recognition, face recognition algorithms [1,2] have achieved impressive results on verification and recognition tasks. The state-of-the-art method, ArcFace [2] achieved a face verification accuracy of 99.83% on LFW [16], which is even better than human-level performance. Another popular task is person re-identification [3,4], which aims at recognizing pedestrians across cameras within a relatively short period, where visual cues are likely to remain consistent. However, these methods are highly sensitive to environmental conditions and inadequate to handle the variations in social media photos or movies, where the faces and bodies are always invisible or blurred and the clothing may be changed.

2.2 Person Recognition in Photo Albums

Person recognition under unconstrained settings is the problem of interest in this work, which mainly focuses on the persons in photo albums. Zhang *et al.* [8] introduced the People In Photo Albums (PIPA) dataset and combined three

visual recognizers on face, body, and poselet-level cues to recognize the persons. To further improve the performance on PIPA, some studies [6,7] paid attention to exploiting more visual cues, such as head [6,7], upper body [6,7], scene [6], pose [7], or other human attributes [6] respectively. Other studies [9,10] focused on combining visual cues and social context to exploit domain-specific information. Li *et al.* [9] exploited contextual cues at person, photo, and group levels and combined them with a heuristic rule to identify persons. Huang *et al.* [10] proposed a framework to couple social context learning with people recognition by a unified formulation, which integrated multiple visual cues adaptively and achieved state-of-the-art performance. However, simply applying image-based methods to recognize persons in videos would lose temporal information and require high computing cost.

2.3 Person Recognition in Unconstrained Videos

Person recognition in unconstrained videos attracts far less attention due to challenges as follows: (1) Lacking annotated video datasets. (2) The temporal and multi-model information of videos put forward higher requirements for algorithms. In particular, the Celebrity Video Identification Challenge [11] held in 2019 presented iQIYI-VID-2019, a large-scale video dataset for multi-modal person recognition. However, almost all (99.65%) videos in this dataset contain clear faces, which is much different from the real-world scenarios. The winning team [12] relied only on face features and achieved better results than others that combining multiple cues [14,15]. Moreover, most teams chose averaging [12] or heuristic rules [13,15] to aggregate the image features of a video in the competition, which are obviously over-simplified. Another noteworthy task is person search in videos. Huang *et al.* [17] proposed a framework to incorporate both the visual similarity and the identity invariance along a tracklet, and developed a new schedule to improve the reliability of propagation, which outperformed mainstream person re-id methods on this problem. However, this task is essentially different, where a clear portrait for each person is required.

3 The Proposed Framework

3.1 Overview

Given a video clip, the task of person recognition in videos is to recognize the identity in the clip, which is defined as a standard supervised classification task [8,11] that we train and test on the same set of identities.

In this work, we devise a Multi-Cue and Temporal Attention framework for this task. As shown in Fig. 2, the overall architecture of MCTA framework mainly contains four parts:

1) Multi-cue region detector and feature extractor. The framework takes as input one video clip which is sampled into F frames. To obtain regions of multiple visual cues, different body parts, including face, head, upper body, and

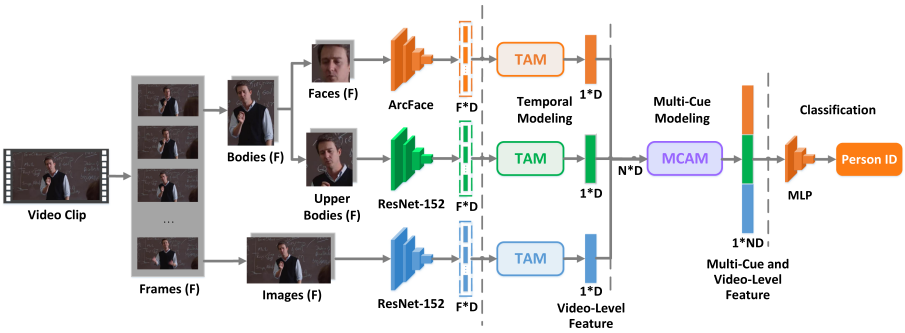

Fig. 2. The overall architecture of multi-cue and temporal attention framework.

the whole body are cropped from frames with region-specific detectors. Next, specific CNNs are adopted to extract the spatial features of these body regions and the whole images.

2) Temporal feature modeling. For each visual cue region, this stage computes a quality score vector for the feature of each sampled frame and aggregates them into a video-level representation. For this, a Temporal Attention Module (TAM) is adopted to adaptively learn the importance of different frames and integrate them.

3) Spatial feature modeling. This stage fuses the video features of different visual cues and forms the final video-level representation for each video clip. For this, a Multi-Cue Attention Module (MCAM) is learned to combine the visual cues from different regions with adaptive weights.

4) Classifier with video-level features. In this stage, the identity of person is predicted by the video-level feature. A three-layer Multilayer Perceptron (MLP) model is adopted for the final classification.

3.2 Temporal Attention Module

Unlike the task of action recognition in videos [18], the movement between frames has little effect on person recognition. For temporal information, we mainly consider the quality of video frames, which is represented by the quality of visual features here. To this end, we devise a Temporal Attention Module (TAM) as the quality predictor, which is inspired by [3], to compute the scores of different frames and aggregate them better.

For each visual cue region, the input of TAM is a feature matrix $\mathbf{X} \in \mathbb{R}^{F \times D}$, where F is the sampled frame number for a video clip and D is the length of feature vector. Then $\mathbf{X}$ is fed into a fully connected layer and a softmax layer, to get the quality score matrix $\mathbf{Z}$:

$$\mathbf{Y} = \mathbf{W}_F \mathbf{X} + \mathbf{b}\,, \tag{1}$$

$$\mathbf{z}_i = \frac{\exp(\mathbf{y}_i)}{\sum_{i=1}^{F} \exp(\mathbf{y}_i)}\,, \tag{2}$$

where $\mathbf{W}_F \in \mathbb{R}^{F\times F}$, $\mathbf{b} \in \mathbb{R}^F$ are parameters to be learned, and $\mathbf{y}_i$ denotes the i-th row of $\mathbf{Y}$. The i-th row of $\mathbf{Z}$, denoted by $\mathbf{z}_i$, is the quality score vector for the i-th frame. Finally, the output feature vector $\mathbf{o}$ is fused through:

$$\mathbf{o} = \sum_{i=1}^{F} \mathbf{z}_i \odot \mathbf{x}_i \,, \tag{3}$$

where $\odot$ is the element-wise multiplication operator, and $\mathbf{x}_i$ denotes the feature vector of the i-th frame. As a result, TAM calculates an attention weighted average on the sequence of image features, which aggregates complementary information from all frames in a video, and the influence of image regions with poor quality is compensated by other frames.

3.3 Multi-Cue Attention Module

Previous studies [6–10] have proved the utility of incorporating multiple visual cues for person recognition. The facial features play a decisive role when the front face is clear. However, in unconstrained videos, the face may be invisible due to the limited scope of camera or occlusion, where we have to resort to other visual cues like hairstyle, clothing, or scene to recognize its identity. Moreover, different visual cue regions always vary in contributions across instances. Inspired by [10], we utilize a Multi-Cue Attention Module (MCAM) to combine the features of multiple visual cues adaptively.

The MCAM is a neural network that takes the stacked features $\mathbf{X} \in \mathbb{R}^{N\times D}$ from all N visual cue regions as input, where N is the number of visual cues and D is the length of feature vector. N positive coefficients are yielded as the weights of different regions through a fully connected layer, a mean operation, and a softmax layer in MCAM:

$$\mathbf{Y} = \mathbf{W}_N \mathbf{X} + \mathbf{b} \,, \tag{4}$$

$$\bar{y_i} = \frac{1}{D} \sum_{j=1}^{D} y_{i,j} \,, \tag{5}$$

$$z_i = \frac{\exp(\bar{y_i})}{\sum_{i=1}^{N} \exp(\bar{y_i})} \,, \tag{6}$$

where $\mathbf{W}_N \in \mathbb{R}^{N\times N}$, $\mathbf{b} \in \mathbb{R}^N$ are learnable parameters, $\bar{y_i}$ denotes the average value of the i-th row $(y_{i,1}, y_{i,2}, ..., y_{i,D})$ of $\mathbf{Y}$, and z_i is the weight for the i-th visual cue region. Then the final feature vector of the i-th visual cue region is given by:

$$\mathbf{o}_i = z_i \mathbf{x}_i \,, \tag{7}$$

where the i-th row of $\mathbf{X}$, denoted by $\mathbf{x}_i$, is the raw feature vector of the i-th visual cue region. Finally, we get the output feature vector $\mathbf{o}$ by the concatenation of $\mathbf{o}_i$. As a result, MCAM can calculate a weighted score for each visual cue adaptively and combine them by weighted concatenation.

4 Experiments

4.1 New Dataset: Character Recognition in Videos

Existing datasets for person recognition under unconstrained settings are mainly based on still images [8,10], while the video-based dataset is rare. To our best knowledge, the latest video dataset for person identification is iQIYI-VID [11], which contains about 200,000 videos of 10,034 identities collected from real online videos. However, as discussed in Sect. 1, almost all video clips in the dataset have clear appearances of faces, which makes it different from the practical scenarios and limited in use.

To facilitate related research and assess our method under more practical settings, we build a high-quality Character Recognition in Videos dataset from movies and TV series, dubbed as CRV. This dataset contains 4,405 video clips of 79 characters in total, which are played by 37 actors or actresses in 34 movies and TV series. The duration of video clips is in the range of 1 to 30 s, with 5.14 s on average. Particularly, all actors and actresses in CRV play multiple (from 2 to 4) roles with differences in one movie or TV series. We need to combine facial features with different visual cues, such as hairstyles, clothing, actions, or surrounding scenes, to identify the specific character in a video clip.

The construction process contains four steps as follows: (1) We first select more than 50 movies and TV series in which one person (or twins) plays multiple roles. Videos of characters acted by one person that are nearly indistinguishable by annotators are excluded. (2) We then ask ten annotators to segment video clips of the selected characters from the movies and TV series. The length of each clip is limited to 1 to 30 s to avoid redundant information while keeping the temporal information. Particularly, each video clip must contain only one or one main person. The scene in one clip should be fixed. Characters with less than 5 clips are discarded. (3) Each candidate video clip is labeled according to its character by another annotator and then checked twice to guarantee the accuracy of both the segmentation and the label. (4) At last, the dataset is randomly split into training, validation, and testing sets by the ratio of 4: 3: 3.

Figure 3 shows several examples of video frames in our dataset. It can be seen that recognizing characters, especially those played by the same actor or actress, is very challenging. Characters in CRV are diverse in views, poses, clothing, and scenes.

4.2 Implementation Details

Data Preprocessing and Feature Extraction. In the experiments, each video clip is uniformly sampled into 16 frames by time. We use three visual cue regions of each frame: face, upper body, and the whole image. For the sampled frames, the bodies of persons are first detected and cropped with Mask R-CNN [19] pre-trained on MS-COCO [20]. Then we separately adopt MTCNN [21] for face detection and alignment, and SSD [22,23] pre-trained on Hollywood-Heads [24] to detect the heads inside the cropped body regions. The regions

Fig. 3. Examples of CRV dataset.

of upper bodies are obtained by simple geometric rules based on the region of heads and bodies. For feature extraction, we adopt ArcFace [2] pre-trained on MS1M-ArcFace [25] to extract face features, and ResNet-152 [26] pre-trained on ImageNet [27] as the feature extractor for other visual cue regions. The face features are duplicated four times and concatenated to have the same dimension (2,048-D) as the other ResNet [26] features. Each feature is then scaled by its maximum absolute value. In particular, we use zero vectors as the features for visual cue regions that are invisible or not detected.

Network Training. In our framework, the features of frames for multiple visual cues are extracted firstly, and the other parts of MCTA are trained end-to-end. A three-layer MLP with a hidden layer of 2,048 nodes is chosen for the final classification based on the validation set. We train MCTA with the cross-entropy loss at a learning rate of 0.0003. The MCTA is implemented based on the Keras framework.

4.3 Comparison with the State-of-the-Art Methods

As discussed in Sect. 2.3, person recognition in videos especially under unconstrained scenarios is a relatively new task, which attracts less attention and is far

Table 1. Comparison to the existing state-of-the-art and baseline methods on CRV.

Method	Module		Result(%)		
	Multi-cue modeling	Temporal modeling	mAP@100	mAP	Accuracy
Face + MLP [12]	–	Average	81.82	83.26	79.98
Multi + MLP	Concatenation	Average	85.71	87.14	87.85
MCTA-t	Concatenation	TAM	86.49	87.90	90.31
MCTA-mc	MCAM	Average pooling	86.63	88.04	90.44
MCTA	MCAM	TAM	**87.01**	**88.42**	**90.56**

from well being solved now. To validate the effectiveness of the proposed MCTA framework, we compare it with existing state-of-the-art and baseline methods on the CRV dataset. The mean Average Precision (mAP) and accuracy are calculated as the evaluation metrics. The details of these methods are as follows:

1) Face + MLP [12]**.** This method uses only face features [2] of frames as the input and adopts an MLP for classification. The averaged probability vector of frames in a video is used as its prediction result. The main idea is similar to [12], the winning team in [11], while we remove tricks like data augmentation and model ensemble. Therefore we consider this model as the state-of-the-art method for person recognition in videos.

2) Multi-Cue + MLP. This method uses the same framework with Face + MLP except that the concatenation of multi-cue features is taken as the input.

3) MCTA-t. This method is a simplified version of MCTA, which replaces the MCAM with simple concatenation to integrate the multi-cue features.

4) MCTA-mc. This method is another simplified version of MCTA, which replaces the TAM with average pooling to combine the features of different frames.

5) MCTA. This is the complete MCTA framework as described in Sect. 3, which adopts the MCAM to integrate features of multiple visual cues and the TAM to model the importance of different frames.

Analysis and Discussion. The results of these methods are listed in Table 1. We can observe that recognizing characters in CRV dataset is challenging. Compared with Face + MLP [12], Multi-Cue + MLP raises the performance by a considerable margin, which validates the significance of multiple visual cues. Moreover, by comparison of Multi-Cue + MLP, MCTA-t, MCTA-mc, and MCTA, we can find that the multi-cue and temporal attention module are both effective in this task. MCAM can generate adaptive weights for different visual cues, and as a result, the mAP is significantly improved. Although the temporal information between frames is poor in a short video clip, TAM gives slightly better performance than average pooling by exploiting the image quality of different frames. Furthermore, MCTA obtains the best performance on CRV, which demonstrates the complementary effect of multiple visual cues and temporal information in videos.

Table 2. Results (%) of different visual cues.

Cue	mAP	accuracy
Face (F)	83.18	86.68
Upper body (U)	72.37	80.09
Image (I)	63.81	74.79
F + U	86.26	89.85
F + I	87.30	89.38
U + I	75.59	82.16
F + U + I	**88.42**	90.56
F + Head + U + Body + I	88.35	**91.40**

4.4 Ablation Study

Significance of Multiple Visual Cues. Here we show the comparison of different visual cues. The results are listed in Table 2. For single visual cue, face feature achieves the highest mAP, which can be used to recognize the person in videos to a pair or group of characters and further recognize it by extra facial information such as expression and decoration. The combination of multiple visual cues can significantly improve the performance, which demonstrates the complementary effect of multi-cue information. Moreover, when we add the features of head and the whole body to MCTA, the performance changes insignificantly. The head feature can provide extra information when the face is invisible, while it's already covered by the upper body feature. Compared with upper body, the whole body always changes in views for different videos, which makes the information unclear.

Significance of Different Modules. Here we explore the effect of multi-cue modeling module, temporal modeling module, and MLP module in MCTA. Table 3 lists the results of the cross combination of different multi-cue information modeling methods, *i.e.*, simple concatenation, heuristic weighted concatenation (the weight ratio is set to F: U: I = 0.4: 0.3: 0.3 after experimental comparison) and MCAM, and different temporal information modeling methods, *i.e.*, average pooling, max pooling, LSTM, and TAM. From the results, we can find that MCAM and TAM perform better than all of the other methods, which proves that our MCTA framework can better explore the spatial-temporal characteristics of persons in unconstrained videos. Moreover, the mAP reduces from 88.42% to 73.92% when we remove the MLP module in MCTA, which reflects the necessity of the MLP part.

Analysis on Hyper Parameters. For the sampled frame number F, we compare the results of MCTA for $F = 1$, 4, 8, 16, and 24. The results are listed in Table 4. It shows that the mAP increases with the growth of input frames (from 71.72% to 88.42%) and reaches a smooth state when F is larger than 16 (88.42%

Table 3. Results (mAP, %) of the cross combination of different multi-cue and temporal information modeling methods.

	Concatenation	Heuristics	MCAM
Average pooling	87.52	87.81	88.04
Max pooling	86.09	86.16	86.18
LSTM	87.00	87.29	87.42
TAM	87.90	88.14	**88.42**

Table 4. Results (%) of different sampled frame number (F).

F	1	4	8	16	24
mAP	71.72	87.60	88.14	88.42	**88.50**

vs. 88.50%). Therefore, we finally choose $F = 16$ for the balance of accuracy and model complexity.

5 Conclusion

In this paper, we propose a new framework named MCTA for person recognition in videos, which utilizes an MCAM to adaptively combine the features of multiple visual cues and a TAM to aggregate the frame-level features by assessing the importance of different frames. We construct a novel dataset CRV from movies and TV series for the recognition of characters under challenging settings. Extensive comparing experiments and ablation studies on CRV show that our approach can learn a better spatial-temporal representation for person recognition in unconstrained videos.

References

1. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR, pp. 815–823 (2015)
2. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR, pp. 4690–4699 (2019)
3. Song, G., Leng, B., Liu, Y., Hetang, C., Cai, S.: Region-based quality estimation network for large-scale person re-identification. In: AAAI, pp. 7347–7354 (2018)
4. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: CVPR, pp. 2138–2147 (2019)
5. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)
6. Oh, S.J., Benenson, R., Fritz, M., Schiele, B.: Person recognition in personal photo collections. In: ICCV, pp. 3862–3870 (2015)
7. Kumar, V., Namboodiri, A., Paluri, M., Jawahar, C.V.: Pose-aware person recognition. In: CVPR, pp. 6223–6232 (2017)

8. Zhang, N., Paluri, M., Taigman, Y., Fergus, R., Bourdev, L.: Beyond frontal faces: Improving person recognition using multiple cues. In: CVPR, pp. 4804–4813 (2015)
9. Li, H., Brandt, J., Lin, Z., Shen, X., Hua, G.: A multi-level contextual model for person recognition in photo albums. In: CVPR, pp. 1297–1305 (2016)
10. Huang, Q., Xiong, Y., Lin, D.: Unifying identification and context learning for person recognition. In: CVPR, pp. 2217–2225 (2018)
11. Liu, Y., et al.: iQIYI celebrity video identification challenge. In: ACM MM, pp. 2516–2520 (2019)
12. Huang, Z., Chang, Y., Chen, W., Shen, Q., Liao, J.: Residualdensenetwork: a simple approach for video person identification. In: ACM MM, pp. 2521–2525 (2019)
13. Fang, X., Zou, Y.: Make the best of face clues in iQIYI celebrity video identification challenge 2019. In: ACM MM, pp. 2526–2530 (2019)
14. Dong, C., Gu, Z., Huang, Z., Ji, W., Huo, J., Gao, Y.: Deepmef: a deep model ensemble framework for video based multi-modal person identification. In: ACM MM, pp. 2531–2534 (2019)
15. Chen, J., Yang, L., Xu, Y., Huo, J., Shi, Y., Gao, Y.: A novel deep multi-modal feature fusion method for celebrity video identification. In: ACM MM, pp. 2535–2538 (2019)
16. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, pp. 1–14 (2008)
17. Huang, Q., Liu, W., Lin, D.: Person search in videos with one portrait through visual and temporal links. In: ECCV, pp. 425–441 (2018)
18. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V.: Temporal segment networks: towards good practices for deep action recognition. In: ECCV, pp. 20–36 (2016)
19. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV, pp. 2961–2969 (2017)
20. Lin, T.Y., et al.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
21. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Sig. Process. Lett. **23**(10), 1499–1503 (2016)
22. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
23. Marin-Jimenez, M.J., Kalogeiton, V., Medina-Suarez, P., Zisserman, A.: LAEO-Net: revisiting people looking at each other in videos. In: CVPR, pp. 3477–3485 (2019)
24. Vu, T.H., Osokin, A., Laptev, I.: Context-aware cnns for person head detection. In: ICCV, pp. 2893–2901 (2015)
25. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 87–102. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_6
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
27. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR, pp. 248–255 (2009)