IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 May 2022, accepted 18 July 2022, date of publication 25 July 2022, date of current version 3 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3193787

RESEARCH ARTICLE

Context-Aware Contribution Estimation for Feature Aggregation in Video Face Recognition

MENG ZHANG[1,2], RUJIE LIU[2], DAISUKE DEGUCHI[1], (Member, IEEE), AND HIROSHI MURASE[1], (Life Fellow, IEEE)

[1]Graduate School of Informatics, Nagoya University, Nagoya 464-8601, Japan
[2]Fujitsu Research and Development Center, Beijing 100022, China

Corresponding author: Meng Zhang (zhang.meng@f.mbox.nagoya-u.ac.jp)

ABSTRACT The difficulties in video-based face recognition, such as dramatic pose variations and low quality, can be alleviated by leveraging the rich complementary information between the frames. However, limited by the mini-batch training strategy, the current deep learning methods only utilizes the frames in each batch during training, which ignore the context of the entire video. In this paper, we propose a context-aware feature aggregation scheme to aggregate complementary information between different frames. Firstly, a two-branch structure is designed as the **C**ontext-aware feature **A**ggregation **N**etwork (**CAN**). Secondly, a context-aware training strategy using a context bank is proposed, which alleviates the limitation of mini-batch samples by using the context of the entire video or several images belonging to the same ID and thus achieves global contribution estimation result. Comparative studies on benchmark datasets, such as IJB-C, YouTube Face (YTF), PaSC and COX, confirm that the proposed approach can achieve state-of-the-art level. Meanwhile, qualitative analysis on Multi-PIE dataset indicates that the contribution learned by the CAN is reasonable and beneficial to video face recognition.

INDEX TERMS Contribution estimation, feature aggregation, video-based face recognition.

I. INTRODUCTION

Video-based face recognition has received increasing interest in both academia and industry, and has been widely used in applications such as security authentication and video surveillance. Although considerable progress has been achieved in still image face recognition owning to the emergence of large-scale datasets and many deep learning-based approaches [1]–[9], video-based face recognition still remains as a significant research challenge. Different from still image face recognition, video face often suffers from low quality, dramatic pose variations, occlusion, and so on. On the contrary, abundant temporal and multi-view information usually exists in the video, which may bring potential to boost accuracy in video face recognition.

To efficiently use more discriminative information in the video, aggregation-based methods [10]–[15] have been widely adopted and impressive performance is gained in video face recognition. The basic idea of aggregation approach is to extract frame-level features at each frame, and then to aggregate them across all frames to form a video-level feature. The most commonly used aggregation technique is average pooling [16], where features of all frames are simply combined with equal importance. However, the low-quality frames would deteriorate the quality of features, resulting in degraded performance of face recognition. Another aggregation method is max pooling [17], which only uses the best quality frame feature as video feature, however, the discriminative information contained in low-quality frames is ignored which can be complementary to high-quality frames.

Recent advance has witnessed deep learning network as an adaptive weighting scheme to aggregate all frame-level features together to form a compact and discriminative video-level feature. Neural aggregation network (NAN) was proposed in [15] for feature combination. It has two modules: one is the CNN feature embedding module to extract the feature representation of each face frame; another is the neural aggregation module to aggregate the video-level feature from

The associate editor coordinating the review of this manuscript and approving it for publication was Shovan Barma.

face video using two attention blocks. QAN [13] adopted two branches scheme, where the one branch is used to extract face feature of each image and the other branch is adopted to predict the quality score of each image. Then, the the final set-level features are obtained by aggregating the features and the quality scores of all images in a set. C-FAN [11] was proposed to learn the quality score of each feature component by adding an aggregation module to the base network, and then to gain the video-level face feature in a video using a single vector aggregated from deep feature vectors. However, limited by the mini-batch training strategy, the quality prediction in the above methods only utilize video frames in each batch during training, which ignore the context of the entire video as well as all frames corresponding to the subject, thus leading to a biased face quality estimation.

We propose a novelty feature aggregation method for video-based face recognition by considering the context of the entire video. Firstly, a context-aware feature aggregation network (CAN) was designed to learn the contribution for each frame in a video, in which the features coming from multiple frames are adaptively aggregated into a compact video-level feature. The network is composed of two branches, one is a feature extractor to extract face feature from a single frame and the other branch is a contribution estimator to estimate the image contribution. The video feature is then aggregated by the features and contributions of all frames in a video clip. Secondly, a context-aware training strategy using a context bank is proposed, where not only the samples in each mini-batch but also the context of entire video clip are considered, thus achieves a global contribution estimation scheme. In addition, in order to reduce the influence of long tail problem in the training corpus, i.e., DeepGlint and Glint360K, a balanced batch selection strategy is further carefully designed. The qualitative analysis on the Multi-PIE dataset shows that the contribution learned by the CAN is reasonable in that it is closely related to image quality, and the quantitative experiments on benchmark datasets indicate that the proposed CAN has achieved the state-of-the-art level.

The main contribution of this paper can be summarized into three folds:

(1) CAN was proposed to learn the contribution for each frame in a video, and the features from multiple frames are adaptively aggregated into a compact video-level feature based on their contributions;

(2) A context-aware training strategy was proposed to achieve global contribution estimation scheme by leveraging the context of entire video clip using a context bank;

(3) A balanced batch selection strategy was carefully designed to reduce the negative impact of the long-tail dataset on performance;

The paper is organized as follows. Section 2 briefly reviews related work on video-based face recognition and feature aggregation methods. Section 3 describes the proposed CAN framework and the contribution-aware training strategy. Finally, we present experimental results in section 4 and discussion in Section 5.

II. RELATED WORK

A. VIDEO-BASED FACE RECOGNITION

Video based face recognition (VFR) has the disadvantages of low resolution and dramatic pose variations compare with still image based face recognition; on the contrary, it also has the advantages of complementary information in consecutive frames. The existing works on VFR can be categorized into two main categories: one is to exploit complementary information contained in multiple video frames, while the other one is to extract higher quality features from each frame.

With frame sequence as input, the person-specific facial dynamics was extracted from continuous video frames using robust face trackers [18], [19]. The aggregation-based methods [11], [13]–[15], [19], [20] were aimed to obtain a compact and discriminative feature aggregated by all frame-level features in a video using adaptive weighting scheme. The key frames selection methods were attempted to gain only a subset of best-quality frames from video clips using frame quality evaluation for efficient face recognition.

Recent deep learning (DL) methods, such as A-softmax [13], CosFace [6], ArcFace [2], etc., introduce margin into the softmax loss to extract more discriminative face features. To solve the blurring problem in video caused by the relative motion between the cameras and the subjects, deblur-based methods [19] deblur the blurred image by estimating a blur kernel, and then to extract the feature. Data uncertainty modeling is another strategy in unconstrained face recognition [21]–[23], especially for noisy images. In their works, data uncertainty learning is applied to capture both the feature (mean) and uncertainty (variance) simultaneously.

B. FEATURE AGGREGATION

The video provides us with abundant complementary information across frames compared with a single image. Therefore, how to aggregate information across frames to get more valuable and effective video-level feature is a crucial issue for robust recognition against variations. Neural Aggregation Network (NAN) [15] proposed two attention blocks to adaptively weight the frames. Discriminative aggregation network (DAN) [24] proposed a network to aggregate raw video frames directly instead of the features obtained by complex processing. Quality aware network (QAN) [13] automatically estimated the quality score for each sample in a set by quality estimator, and weighted all frames by the predicted quality scores. Region-based Quality Estimation Network (QAN+) [20] further extended the idea of QAN into local regions, which used an ingenious training mechanism to extract the complementary region-based information between different frames. COmpact second-order network (COSONet) [25] proposed a second-order network to extract features from faces with large variations and a mixture loss function to encourage the discriminate and simultaneously regularizes the feature. Multicolumn Network (MN) [14] taken entire images in a set as input, and learned to compute a set-level fix-sized feature representation.

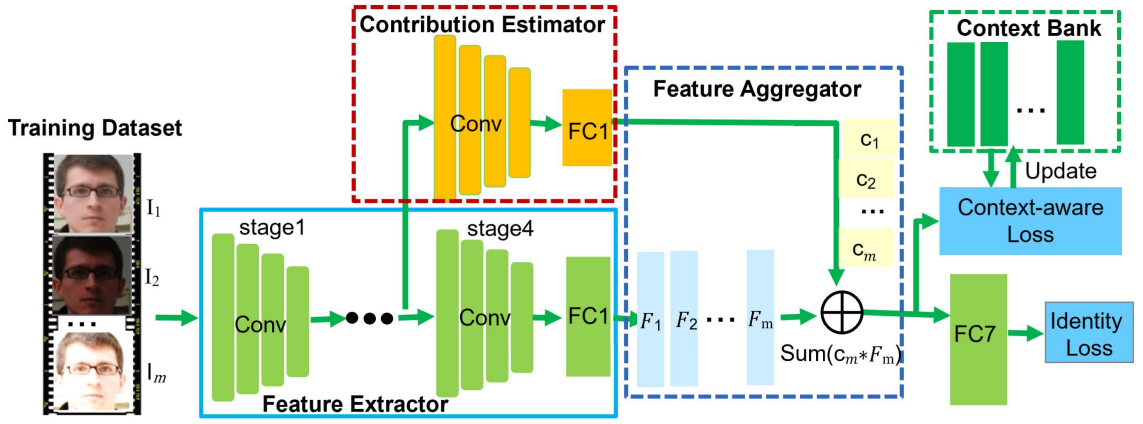


FIGURE 1. Architecture of the CAN. The input of the CAN is the video clip or several images belonging to the same ID. The feature extractor is a base model, which is used to extract each frame feature of the video clip. The contribution estimator is added to the based model using several convolution layers and one-node fully connected layer, which is used to estimate the contribution of each frame to their video clip. The feature aggregator is used to aggregate the contribution scores and features of all frame in the video clip. The final video-like feature is thus directly obtained by feature aggregator. The context bank is maintained to memorize the globally features. Here, we use video-level identity loss and context-aware loss to supervise the training. (Best viewed in colors.)

Each component of the feature vector may encode different subsets of facial features, thus bias could be caused when we emphasize or suppress all components simultaneously. To alleviate this problem, a meta attention-based aggregation scheme is used in [19], to adaptively fine-grain the weights along each feature dimension among all frames so as to handle the feature on dimension level. Similarly, component-wise feature aggregation scheme is used in C-FAN [11], for video face recognition, where the quality value for each feature component is separately learned. C-FAN automatically learns to suppressing features with low-quality scores, while retain salient face features with high-quality scores.

As a summary, the aim of feature aggregation methods is to automatically learn the weights from frame level or feature component level, and the quality criterion is used therein to represent the importance of each single frame or each feature component, therefore, these methods are usually called as quality based feature aggregation methods.

However, limited by the mini-batch training strategy, the existed quality-based feature aggregation methods failed to globally consider the relation among frames in a video clip or all samples of one ID during training, thus lead to bias or inaccuracy in quality estimation. This motivates us to seek a better solution in this paper, especially to investigate valuable information in the low-quality images.

III. PROPOSED APPROACH

In this section, we first describe the CAN, which incorporates feature extractor network and contribution estimator network to obtain respectively the feature and the contribution of a single frame. Then, the context-aware training strategy is introduced, where not only the samples in each mini-batch but also the context of entire video are considered.

A. CAN

The CAN network architecture consists of three modules: feature extractor, contribution estimator, and feature aggregator, as illustrated in Fig.1.

1) FEATURE EXTRACTOR

most popular deep neural networks (e.g., ResNet34 in ArcFace [2]) can be adopted as backbone to extract the feature from each frame [26], [27]. Once built, the extractor is kept fixed during the training of the CAN network.

2) CONTRIBUTION ESTIMATOR

a new branch is adopted as the contribution estimator by connecting the feature extractor, aiming to get the contribution of each single frame, which is then used as the weight in later feature aggregation stage. The structure may have different complexity, ranging from one-node fully connected layer to one or several convolution layers. More complex network may bring higher learning ability, but certainly with the cost of extra computation and the risk of over-fitting. From the experimental results in section 4, we have found that the contribution value is highly correlated with image quality, the better the quality of a frame the higher the contribution value. That means, the estimated contribution is in fact a physically meaningful factor.

3) FEATURE AGGREGATOR

Assume $\mathrm{V} = \{I_1, I_2, \ldots, I_m\}$ be a video clip, with I_i indicating the i-th frame. Let $F(\cdot)$ and $C(\cdot)$ respectively denote the feature extractor and contribution estimator, which outputs the feature vector f_i and attribution value c_i for each frame, i.e., $f_i = F(I_i)$ and $c_i = C(I_i)$. The final face representation of a video is thus directly obtained by weighted average of

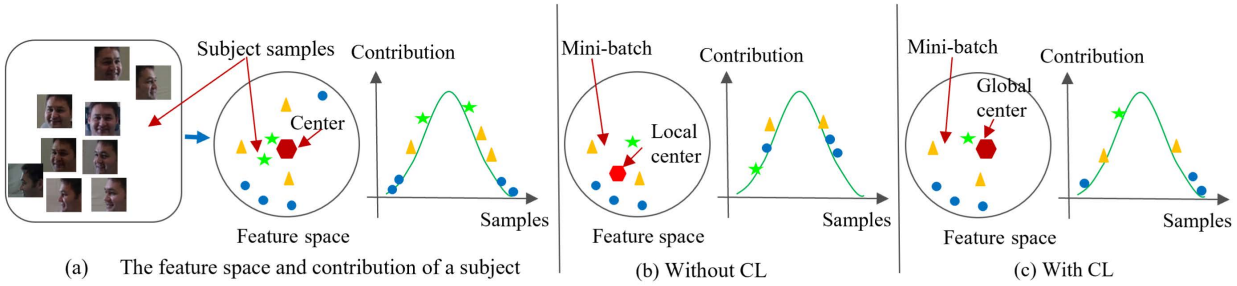


FIGURE 2. Contribution estimation with the proposed context-aware contribution loss (CL). (a) Feature space and contribution value of the samples belonging to a subject. The nearer a sample is to the ID center, the higher (better) the contribution (quality). (b) Feature space and contribution of a mini-batch samples. The samples close to the mini-batch center are assigned higher contribution value even they are far from the ID center. However, the samples close to the ID center but far from the mini-batch center are assigned a low contribution value, such as the estimated contribution value of the sample represented by the pentagram, thus leading to a biased face contribution estimation. (c) Contribution estimation result under the CL loss, where the samples far from ID center are assigned lower contribution value even they are close to the mini-batch center. Meanwhile, the samples close to the ID center are assigned higher contribution value even they are close to the mini-batch center, such as the estimated contribution value of the sample represented by the pentagram. (Best viewed in colors.)

the features of the video frames, as follows:

$$F_V = \emptyset\left(f_1, f_2, \ldots f_m\right) = \frac{\sum_{i=1}^{m} c_i * f_i}{\sum_{i=1}^{m} c_i} \tag{1}$$

Optionally, we may use the frame feature with highest contribution value as the video feature, i.e., feature selection scheme.

B. CONTEXT-AWARE TRAINING

We proposed two loss to train our model, one is video-level identity loss, while another is context-aware contribution loss. With the video-level identity loss, ground-truth of contribution or quality value of each frame is not necessary during training, which largely reduces the cost of building training data in our solution. With the context-aware contribution loss, a context-aware features memory bank is used to store more information (beyond the information in each training batch as in traditional method) during training stage, thus better accuracy can be achieved.

1) VIDEO-LEVEL IDENTITY LOSS

The video-level feature is firstly obtained by the contribution weighted aggregation of the features of all frames in a video clip, and an ArcFace [2]-like loss is chosen to penalize video-level identify error. In such case, no contribution or quality value of each frame is provided as ground-truth for supervision, so, the training can be seemed as unsupervised.

The video-level identify loss is defined as:

$$L_{ID} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{ji}+m))} + \sum_{j=1, j\neq y_i}^{n} e^{s(\cos\theta_j)}} \tag{2}$$

where θ is the angle between a video-level feature and the corresponding weight, N is a number of video clips in a mini-batch, m is a marginal factor, s is a scale factor.

2) CONTEXT-AWARE CONTRIBUTION LOSS (CL)

Given traditional DL model training strategy where the calculation is limited to the mini-batch samples, contribution prediction is achieved only based on the video frames in each mini-batch, while ignoring context information in the entire video or even the whole corpus, as illustrated in Fig.2.

To alleviate the above limitations, we always keep a renewed global representation for each ID using a memory bank, and force the local representation calculated from each mini-batch to be close to the global representation. The global representation is obtained from the entire video (or whole corpus), thus context information is introduced to the mini-batch based training. Let $\mathrm{B} = \left\{F_{g1}, F_{g2}, \ldots, F_{gi}\right\}$ be the memory bank, with F_{g_i} representing the global representation feature of the i-th class. The context-aware loss is thus can be defined as follows:

$$L_C = \sum_{i=0}^{n}\left\|F_{v_i} - F_{gi}\right\| = \sum_{i=0}^{n}\left\|\frac{\sum_{j=0}^{m} c_j * f_j}{\sum_{j=0}^{m} c_j} - F_{gi}\right\| \tag{3}$$

where, F_{v_i} is the video-level feature as introduced in 3.1, F_{gi} is the global video-level feature of class which is got by simply averaging all the video-level features belonging to i-th class. The F_{gi} are updated by calculating the average of the features of the same class in the mini-batch on each iteration. If the mini-batch doesn't include the set features of some classes, their corresponding global features will be not updated.

The gradients of L_C with respect to F_{v_i} and F_{gi} are given by:

$$\frac{\partial L_C}{\partial F_V} = \frac{1}{n}\left(F_{v_i} - F_{gi}\right) \tag{4}$$

$$\Delta F_{gi} = \frac{\sum_{i=1}^{n}\delta\left(y_i = j\right)\cdot\left(F_{gi} - F_{v_i}\right)}{\varepsilon + \sum_{i=1}^{n}\delta\left(y_i = j\right)} \tag{5}$$

where, δ is a Kronecker Delta function. ε is a small positive number to avoid zero denominator, which is set to 10^{-5}.

3) COMBINATION OF THE TWO LOSSES

Finally, the video-level identity loss and contribution loss are combined to jointly train our model, as follows:

$$L = L_{ID} + \lambda L_C \tag{6}$$

where, λ is adopted for balancing the two loss functions.

IV. EXPERIMENT AND RESULTS

In this section, several commonly used benchmark datasets are firstly introduced. Then, the implementation details of the proposed algorithm are represented. To gain more insight into how our model behaves, both qualitative and quantitative analysis as well as ablation study are presented. Besides this, we compare our proposed approach with state-of-the-art approaches to confirm its effectiveness.

A. DATASETS

Both feature extractor and the contribution estimator are trained on DeepGlint and Glint360k dataset, while the accuracy is evaluated on three benchmarks, i.e., COX, IJB-C, PaSC and COX. In addition, Multi-PIE dataset is used for qualitative illustration of the effect of contribution estimation.

DeepGlint includes cleaned MsCeleb1M [2], and Asian celebrity dataset with totally 6.6M celebrity images of 172K subjects. Therefore, it's adopted for the analysis of the proposed method and some ablation study with the relatively small data set for training.

Glint360K [28] is a widely adopted large-scale dataset for training face recognition model, which includes more than 17M images from 360K subjects merged from clean Celeb-500K [29] and MS1M-Retinaface [2]. So, to verify the effectiveness of our proposed method on large-scale datasets, Glint360K is used as training set for comparing with SOTA methods.

COX [30] dataset contains 1,000 subjects, including 1 still image and 3 Videos for each subject, a total of 1,000 still images and 3,000 videos. The videos were captured at different locations by three cameras, and each subject was walking in a large gym to simulate the surveillance scenario. Three standard matching protocols were also proposed by the author for face identification testing, i.e., video-to-still (V2S), as shown in Fig. 3.

IJB-C [31] dataset includes 3,531 subjects with a total 117,542 unconstrained video frames from 11,779 videos and 31,334 still facial images, which is an extension of IJB-B. In the 1:1 verification, there are 23,124 templates with 15,639K impostor matches and 19,557 genuine pairs. The verification protocol of IJB-C contains more impostor pairs, thus the True Accept Rates (TAR) at lower False Accept Rates (FAR) is used in here, as shown in Fig. 4.

PaSC [32] dataset contains 265 subjects, and 2,802 videos. Half of its videos are captured by controlled camera (denoted as PaSC-C), while the rest are captured by hand held camera (denoted as PaSC-H), and each subject is asked to do some predefined actions. Therefore, the face photos cover serious video-type noises and large pose variations. Our evaluate on PaSC totally follows the Predefined face verification protocol, as shown in Fig. 5.

YTF [33] dataset was downloaded from YouTube, which includes 3,425 videos of 1,595 subjects with an average of 2.15 videos of each subject. The average length of a video clip is 181.3 frames, of which the longest clip is 6,070 frames and the shortest clip duration is 48 frames, as shown in Fig. 6.

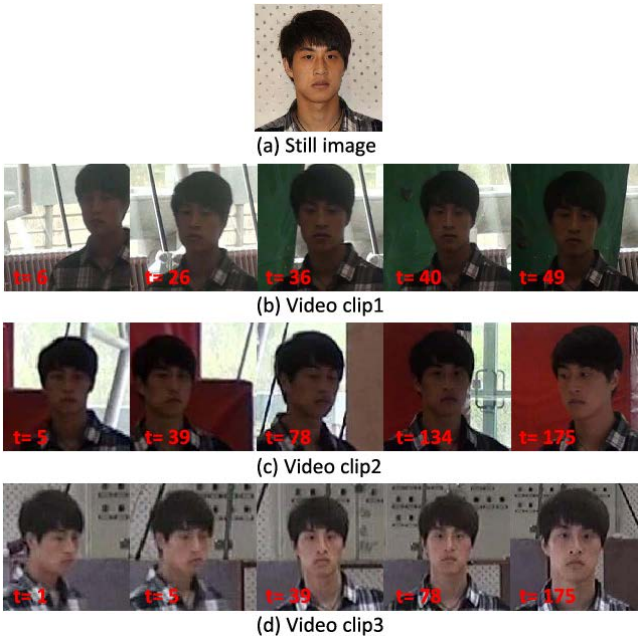


FIGURE 3. Examples of images in the COX. The videos were captured at different locations by three cameras, and each subject was walking in a large gym to simulate the surveillance scenario.

FIGURE 4. Sample images in the IJB-C, which is an extension of the IJB-A dataset with about 138,000 face images, 11,000 face videos, and 10,000 non-face images. All subjects in the dataset are ensured to appear in at least two still images and one video.

Multi-PIE [34]. CMU Multi-PIE provides face images of 337 subjects with comprehensive variety in illumination, expression and pose, by carefully designing the configuration of 15 cameras and 18 flashes, thus it is very suitable to confirm the correlation of contribution estimation with image quality, as shown in Fig. 7.

B. IMPLEMENTATION DETAILS

1) DATA AUGMENTATION (DA)

Blur is imposed to the Deepglint and Glint360K to simulate video-like training data, where one-dimensional local averaging of neighboring pixels is used to generate motion blur, while a Gaussian kernel was adopted to simulate the out-of-focus blur. Besides this, we split the images into 5×5 blocks and randomly replace some blocks with black masks for the augmentation of occlusion data. Illumination variance is

FIGURE 5. Sample images in the PaSC from four sessions. Note that pose, distance to camera and sensor were varied within sessions, while locations were varied between sessions.

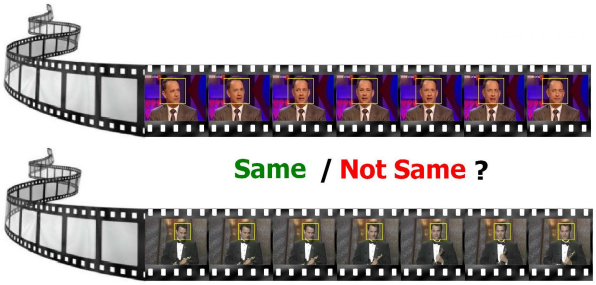


FIGURE 6. Sample images in the YTF downloaded from YouTube, which is designed for studying the problem of unconstrained face recognition in videos.

FIGURE 7. Sample images in the Multi-PIE.

achieved by simply adjusting the brightness of the training images.

2) BALANCE STRATEGY (BS)

Long tail distribution of the training data, the fact that a small number of entities appear frequently while most of others remain relatively rare, usually poses great impact on the feature learning process and feature extraction ability. To solve this problem, two strategies are designed in this paper. One straightforward strategy is to remove the very head and tail subjects. More specifically, the subjects with more than 500 and less than 10 samples are removed. Given that there are enough subjects being remained, this strategy brings some improvement on the accuracy. The second strategy is that we select samples for each mini-batch in training based on ID but not individual images. In image based mini-batch solution, the subjects with more samples will have a higher probability to be selected. However, in ID based solution, we randomly select n IDs from the ID list and then get m images for each ID to generate the mini-batch of $n*m$ samples, thus avoid the bias to head subjects.

TABLE 1. Results on IJB-C dataset with 1:1 verification protocol (TAR@FAR=10^{-3}, 10^{-4}, 10^{-5}). "CAN" mean the proposed context-aware feature aggregation network.

Method	10^{-5}	10^{-4}	10^{-3}
Multicolumn [14]	77.10	86.20	92.70
ArcFace [2]	87.28	92.13	95.55
PFE [22]	89.64	93.25	95.49
DUL [21]	87.22	92.43	95.38
GroupFace [4]	94.53	96.26	-
VPL [4]	-	96.76	-
R100, CAN, DeepGlint	95.44	96.88	97.87
R100, CAN, Glint360k	**96.29**	**97.62**	**98.52**

3) TRAINING PARAMETERS

The ResNet50 (R50) and ResNet100 (R100) are pre-built as the feature extractor of small model and large model respectively. After that, the feature extractor is fixed and the contribution estimator is then trained on the same dataset with the above data augmentation scheme. Stochastic gradient descent (SGD) is used with momentum and weight decay value being 0.9 and 0.005. The value of λ is set to be 0.1, and the size of mini-batch is set to be 100 including 20 randomly selected subjects and 5 images per subject.

C. COMPARISON WITH STATE-OF-THE-ART METHODS

Following the standard evaluation protocols, we compare our proposed model with the state-of-the-art methods on several benchmarks, i.e., IJB-C, PaSC, YTF and COX.

1) EVALUATION ON IJB-C DATASET

The commonly used criterion of true acceptance rate at different false acceptance rate (TAR@FAR=10^{-3}, 10^{-4}, 10^{-5}) is used for the evaluation on IJB-C dataset. We compare our proposed method with several state-of-the-art face recognition methods including both feature aggregation and non-aggregation methods, and tabulate the result in Table 1. For the purpose of fair comparison, both DeepGlint and Glint360k are implemented here. It can be clearly seen from the results that the proposed model outperforms the non-aggregation methods with a large margin, i.e., 8.12% and 9.01% better than ArcFace at FAR=10^{-5} trained on DeepGlint and Glint360k respectively.

2) EVALUATION ON PaSC DATASET

We further test our method in surveillance scene by using PaSC dataset, with the result shown in Table 2. Similar to

TABLE 2. Results on PaSC dataset with 1:1 verification protocol (TAR@FAR=10^{-2}) and YTF dataset (Accuracy(%)). "PaSC-C" mean videos captured by controlled camera. "PaSC-H" mean videos captured by hand held camera. "CAN" mean the proposed context-aware feature aggregation network.

Method	PaSC-C	PaSC-H	YTF
NAN [15]	-	-	95.72
QAN [13]	-	-	96.17
DAN [24]	92.00	80.30	94.28
ADRL [35]	95.67	93.78	96.52
TBE-CNN [36]	96.20	95.80	94.96
COSONet [25]	97.40	96.00	-
C-FAN [11]	-	-	96.50
R100, CAN, DeepGlint	97.67	96.83	97.18
R100, CAN, Glint360k	**98.46**	**97.62**	**97.53**

TABLE 3. Rank-1 identification rates (%) under the V2S setting for different methods on the COX face database. "CAN" mean the proposed context-aware feature aggregation network.

Model	V2S_1	V2S_2	V2S_3
PSCL [30]	38.60 ± 1.39	33.20 ± 1.77	53.26 ± 0.80
LERM [37]	45.71 ± 2.05	42.80 ± 1.86	58.37 ± 3.31
VGG Face [17]	88.36 ± 1.02	80.46 ± 0.76	90.93 ± 1.02
TBE-CNN [36]	93.57 ± 0.65	93.69 ± 0.51	98.96 ± 0.17
R100, CAN, DeepGlint	96.14 ± 0.49	94.69 ± 0.25	99.68 ± 0.09
R100, CAN, Glint360k	$\mathbf{98.21 \pm 0.28}$	$\mathbf{95.18 \pm 0.19}$	$\mathbf{99.86 \pm 0.07}$

that in IJB-C, our method again behaves consistently better than the literature methods. The video content in PaSC, especially the handheld case, suffers from more severe conditions due to camera shaking, pose, blur, etc., therefore, most of the face images in each video clip are of low quality. The simple average method, such as average pooling aggregation, assigns equal weights to each frame. In this way, the low quality frames with improper features would degrade the performance of final recognition. On the contrary, our contribution estimator obtain the contribution value closely related to image quality, thus depressing the low quality frames and strengthening the contribution of high quality frames. Our CAN method outperforms ADRL aggregation by 2.79% at FAR=10^{-2} in control scene, while this superiority increases to be 3.84% at FAR=10^{-2} in handheld scene, which implies the robustness of our method to deteriorated image quality.

3) EVALUATION ON YTF DATASET

The result of a 10-fold cross-validation is calculated on YTF dataset as in Table 2. Compared with other aggregation methods, no video face datasets were used in training our contribution estimator. Even though, better performance is still achieved, which confirms the superiority of our method among these state-of-the-arts, i.e., 1.03% better than C-FAN.

4) EVALUATION ON COX DATASET

The Rank-1 identification rates on COX is listed in Table3. Again, our method outperforms literature methods. Especially on camera 1 and camera 2, more than 4% improvement is achieved.

TABLE 4. Ablation experiment on PaSC dataset with 1:1 verification protocol (TAR@FAR=10^{-2}). "Conv" mean convolution module. "CS" mean context-aware strategy. "DA" mean data augmentation. "BS" mean balance strategy. "PaSC-C" mean videos captured by controlled camera. "PaSC-H" mean videos captured by hand held camera.

Model	Conv	CS	DA	BS	PaSC-C	PaSC-H
Baseline	-	-	-	-	93.43	80.34
A	-	✓	✓	✓	93.41	84.18
B	✓	-	✓	✓	95.79	94.03
C	✓	✓	-	✓	93.39	89.16
D	✓	✓	✓	-	96.15	93.89
E	✓	✓	✓	✓	**97.41**	**97.05**

D. ABLATION STUDIES

In this part, we perform ablation studies to understand the contribution of each technique, ranging from the model such as the structure and loss to the implementation skills such as data augmentation and balance strategy. The ResNet50 with DeepGlint is adopted for the ablation study with the relatively small data set for training. TAR@FAR on PaSC dataset is used here for the comparison and the results are shown in Table 4.

1) MODEL STRUCTURE

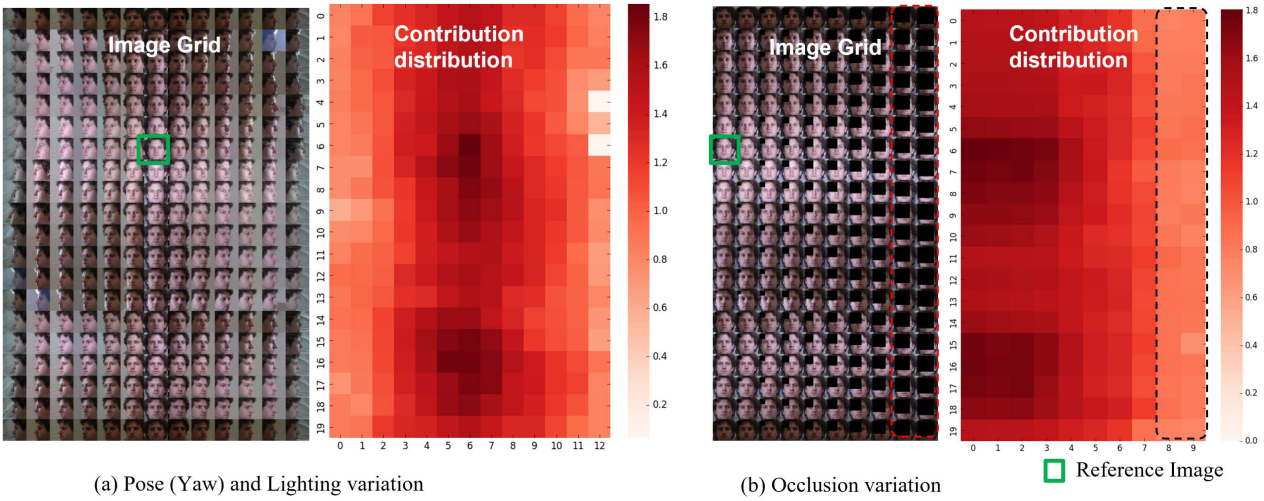
Three contribution estimators of different complexity are firstly compared, (1) baseline structure, where only feature extractor is adopted and the features from all frames are aggregated by average pooling; (2) FC1 (Model A), this is a simplified contribution estimator where only a one-node fully connected module is used; (3) Conv+FC1 (Model E), this is our solution comprising a convolution module and a one-node fully connected module. It can be seen that introducing contribution estimator evidently boosts the recognition accuracy, and this is more obvious in handheld case, e.g., 4% improvement on handheld data. As mentioned early, the images in handheld PaSC suffer severe quality degradation, thus it is even crucial to select the most informative and discriminative frames for correct recognition. However, the structure of one node FC is too simple, thus the learning capability may be not enough. By adding additional convolution layers, the learning capability is enhanced and much bigger improvement can be further obtained, e.g., 12.87% improvement on handheld data.

2) CONTEXT-AWARE STRATEGY (CS)

We compare different learning strategies, i.e., with or without CS (Model E or Model B). It is obvious that CS brings further benefits to the accuracy, especially in handheld case, i.e., 3.02% improvement in comparison not using this strategy. This result once again confirms the effectiveness of our contribution estimation scheme in aggregating low quality video frames for recognition.

3) DATA AUGMENTATION (DA)

Data augmentation plays an important role in our video face recognition solution. The photos in DeepGlint, which are



(a) Pose (Yaw) and Lighting variation (b) Occlusion variation

FIGURE 8. Shows the contribution distribution across varied pose, illumination and occlusion on Multil-PIE. (a) Samples of a subject across varied pose, illumination and corresponding contributions predicted by our contribution estimator. From left to right, faces with different poses are represented, spaced in 15 degree intervals. (b) Samples of a subject with varied occlusion and corresponding contributions predicted by our contribution estimator. (Best viewed in colors.)

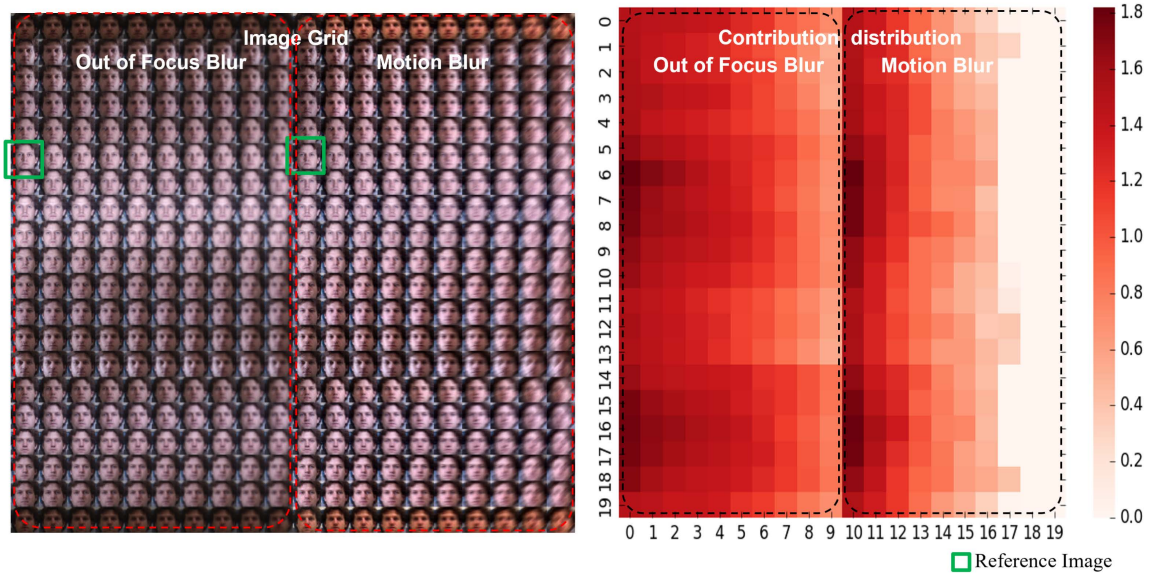


FIGURE 9. Shows the contribution distribution on different motion blur and out of focus blur on Multil-PIE, where the left is an artificially adding motion and out of focus blur to the face image to simulate the different types of blur and the right is the corresponding predicted contribution of the contribution estimator. (Best viewed in colors.)

usually captured under good conditions or even captured by professional photographers, are much different from surveillance scenario, e.g., the photos are usually high resolution and don't have blur. The feature extractor and estimator built on DeepGlint thus can't behave well on PaSC video data. By introducing data augmentation to the training corpus, the data becomes more consistent with the video scene, and better accuracy can be obtained (Model E or Model C). This can be confirmed by the result, where 4.02% and 7.89% improvement is obtained for control and handheld cases respectively.

4) BALANCE STRATEGY (BS)

As introduced in 4.1, long tail problem still exists in the DeepGlint corpus, therefore, The balance strategy may make contribution to the accuracy without any surprise. The balance strategy itself may be not a part of our context-aware feature aggregation algorithm, however, it is a helpful training scheme toward better accuracy. Compared with the Model E and the Model D, it is obvious that BS brings further benefits to the accuracy, i.e., 3.16% improvement for PaSC-H in comparison not using this strategy.

E. QUALITATIVE ANALYSIS

We qualitatively illustrate the result of contribution estimation on Multi-PIE dataset, as shown in Fig. 8 and Fig. 9.

It can be clearly seen that the estimated contribution value is closely related to face pose and illumination from the Fig. 8(a). For example, the frontal face image with normal light has high contribution. With the increase of pose and the decrease of illumination, the contribution value is decreasing, and the images in extreme illumination condition and large pose tend to obtain very low scores, which indicates positive correlation between contribution estimation value and image quality.

To conduct qualitative analysis on the effectiveness of the contribution estimator on occlusion, we artificially add an occlusion to the face image to simulate the different occlusion. The results are shown in Fig. 8(b). It can be seen that by adding an occlusion into the original images, the contribution value will decrease gradually, which prove the effectiveness of contribution estimation for occlusion situations.

To further conduct qualitative analysis on the effectiveness on blur, we also artificially add motion and out of focus blur to the face image to simulate the different types of blur, as is shown in Fig. 9. The image contribution will decrease gradually by adding more blur into the original images, which prove the effectiveness of contribution estimation for blur image.

V. CONCLUSION

In this paper, we propose a new feature aggregation based method for video-based face recognition by considering the context of entire video, i.e., a context-aware feature aggregation scheme to aggregate complementary information between different frames in video. Several innovative points are presented, including: (1) a two-branch DL network for context-aware feature aggregation; (2) a context-aware training strategy for global contribution estimation by utilizing the context of entire video clip using a context bank; (3) a balanced batch selection strategy for better accuracy by reducing the negative impact of long-tail training dataset. While our approach is proposed for video-based face recognition, it can also be performed in other computer vision fields, especially for video-based object recognition problem.

REFERENCES

[1] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* (Lecture Notes in Computer Science), vol. 9911, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 499–515, doi: 10.1007/978-3-319-46478-7_31.

[2] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[3] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "CurricularFace: Adaptive curriculum learning loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5901–5910.

[4] Y. Kim, W. Park, M.-C. Roh, and J. Shin, "GroupFace: Learning latent groups and constructing group-based representations for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5621–5630.

[5] X. Shan, Y. Lu, Q. Li, and Y. Wen, "Model-based transfer learning and sparse coding for partial face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4347–4356, Nov. 2021, doi: 10.1109/TCSVT.2020.3047140.

[6] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[7] S. Yang, W. Deng, M. Wang, J. Du, and J. Hu, "Orthogonality loss: Learning discriminative representations for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2301–2314, Jun. 2021, doi: 10.1109/TCSVT.2020.3021128.

[8] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10823–10832.

[9] M. Zhang, R. Liu, H. Nada, H. Uchida, T. Matsunami, and N. Abe, "A pairwise learning strategy for video-based face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 38–44.

[10] H. Cevikalp, H. S. Yavuz, and B. Triggs, "Face recognition based on videos by using convex hulls," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4481–4495, Dec. 2020, doi: 10.1109/TCSVT.2019.2926165.

[11] S. Gong, Y. Shi, and A. K. Jain, "Video face recognition: Component-wise feature aggregation network (C-FAN)," 2019, *arXiv:1902.07327*.

[12] W. Heng, T. Jiang, and W. Gao, "How to assess the quality of compressed surveillance videos using face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2229–2243, Aug. 2019, doi: 10.1109/TCSVT.2018.2866701.

[13] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5790–5799.

[14] W. Xie and A. Zisserman, "Multicolumn networks for face recognition," 2018, *arXiv:1807.09192*.

[15] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4362–4371.

[16] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition from video," in *Proc. Eur. Conf. Comput. Vis.*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Germany: Springer, 2012, pp. 766–779.

[17] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015. [Online]. Available: https://ora.ox.ac.uk/objects/uuid:a5f2e93f-2768-45bb-8508-74747f85cad1/download_file?file_format=pdf&safe_filename=parkhi15.pdf&type_of_work=Confer

[18] M. Bicego, E. Grosso, and M. Tistarelli, "Person authentication from video of faces: A behavioral and physiological approach using pseudo hierarchical hidden Markov models," in *Advances in Biometrics*, D. Zhang and A. K. Jain, Eds. Berlin, Germany: Springer, 2005, pp. 113–120.

[19] N. Sankaran, S. Tulyakov, S. Setlur, and V. Govindaraju, "Metadata-based feature aggregation network for face recognition," in *Proc. Int. Conf. Biometrics (ICB)*, Gold Coast, QLD, Australia, Feb. 2018, pp. 118–123, doi: 10.1109/ICB2018.2018.00028.

[20] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," 2017, *arXiv:1711.08766*.

[21] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5710–5719.

[22] Y. Shi and A. Jain, "Probabilistic face embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6902–6911.

[23] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain, "Towards universal representation learning for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6817–6826.

[24] Y. Rao, J. Lin, J. Lu, and J. Zhou, "Learning discriminative aggregation network for video-based face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3781–3790.

[25] Y. Mao, R. Wang, S. Shan, and X. Chen, "COSONet: Compact second-order network for video face recognition," in *Proc. Asian Conf. Comput. Vis.*, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Cham: Springer, 2019, pp. 51–67.

[26] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi, and H. Ghayvat, "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, Oct. 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/20/2470

[27] C. Patel, D. Bhatt, U. Sharma, R. Patel, S. Pandya, K. Modi, N. Cholli, A. Patel, U. Bhatt, M. A. Khan, S. Majumdar, M. Zuhair, K. Patel, S. A. Shah, and H. Ghayvat, "DBGC: Dimension-based generic convolution block for object recognition," *Sensors*, vol. 22, no. 5, p. 1780, Feb. 2022.
[28] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, and Y. Fu, "Partial FC: Training 10 million identities on a single machine," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCV)*, Montreal, BC, Canada, Oct. 2021, pp. 1445–1449, doi: 10.1109/ICCVW54120.2021.00166.
[29] J. Cao, Y. Li, and Z. Zhang, "Celeb-500k: A large training dataset for face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 2406–2410, doi: 10.1109/ICIP.2018.8451704.
[30] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen, "A benchmark and comparative study of video-based face recognition on COX face database," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5967–5981, Dec. 2015.
[31] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA Janus benchmark—C: Face dataset and protocol," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 158–165.
[32] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng, "The challenge of face recognition from digital point-and-shoot cameras," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Arlington, VA, USA, Sep. 2013, pp. 1–8 doi: 10.1109/BTAS.2013.6712704.
[33] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, Jun. 2011, pp. 529–534, doi: 10.1109/CVPR.2011.5995566.
[34] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–8.
[35] Y. Rao, J. Lu, and J. Zhou, "Attention-aware deep reinforcement learning for video face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3931–3940.
[36] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 1002–1014, Apr. 2018.
[37] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning Euclidean-to-Riemannian metric for point-to-set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1677–1684.

MENG ZHANG received the B.C. degree from the Qilu University of Technology, China, in 2013, and the M.S. degree from the Beijing University of Technology, China, in 2016. He is currently pursuing the Ph.D. degree in intelligent systems with the Graduate School of Informatics, Nagoya University, Japan. Since then, he has been a Researcher at Fujitsu Research and Development Center Company Ltd., Beijing, China. His research interests include pattern recognition, computer vision, face recognition, and deep learning.

RUJIE LIU received the B.C., M.S., and Ph.D. degrees in electronic engineering from Beijing Jiaotong University, in 1995, 1998, and 2001, respectively. Since then, he worked as a Researcher at Fujitsu Research and Development Center Company Ltd., Beijing, China. He has published more than 40 articles and tens of inventions. His research interests include AI, pattern recognition, and image processing.

DAISUKE DEGUCHI (Member, IEEE) received the B.Eng. and M.Eng. degrees in engineering and the Ph.D. degree in information science from Nagoya University, Japan, in 2001, 2003, and 2006, respectively. He became a Postdoctoral Fellow at Nagoya University, in 2006. From 2008 to 2012, he was an Assistant Professor at the Graduate School of Information Science. From 2012 to 2019, he was an Associate Professor at Information Strategy Office. Since 2020, he has been an Associate Professor with the Graduate School of Informatics. He is working on the object detection, segmentation, recognition from videos, and their applications to ITS technologies, such as detection and recognition of traffic signs. He is a member of IEICE and IPS Japan.

HIROSHI MURASE (Life Fellow, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from Nagoya University, Japan. In 1980, he joined Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993, he was a Visiting Research Scientist with Columbia University, New York. He has been a Professor with Nagoya University, since 2003. His research interests include computer vision, pattern recognition, and multimedia information processing. He is a fellow of the IPSJ and the IEICE. He was awarded the IEEE CVPR Best Paper Award, in 1994, the IEEE ICRA Best Video Award, in 1996, the IEICE Achievement Award, in 2002, the IEEE Multimedia Paper Award, in 2004, and the IEICE Distinguished Achievement and Contributions Award, in 2018. He received the Medal with Purple Ribbon from the Government of Japan, in 2012.