Low Quality Video Face Recognition: Multi-mode Aggregation Recurrent Network (MARN)

Sixue Gong, Yichun Shi, Anil K. Jain
Michigan State University, East Lansing, MI
gongsixu@msu.edu, shiyichu@msu.edu, jain@cse.msu.edu

Abstract

Face recognition performance deteriorates when face images are of very low quality. For low quality video sequences, however, more discriminative features can be obtained by aggregating the information in video frames. We propose a Multi-mode Aggregation Recurrent Network (MARN) for real-world low-quality video face recognition. Unlike existing recurrent networks (RNNs), MARN is robust against overfitting since it learns to aggregate pre-trained embeddings. Compared with quality-aware aggregation methods, MARN utilizes the video context and learns multiple attention vectors adaptively. Empirical results on three video face recognition datasets, IJB-S, YTF, and PaSC show that MARN significantly boosts the performance on the low quality video dataset while achieves comparable results on high quality video datasets.

1. Introduction

An increasing number of videos captured by both mobile devices and CCTV systems around the world[1] has generated an urgent need for robust and accurate face recognition in low quality video. Approaches to face recognition for high quality still images (controlled capture and cooperative subjects) are not able to deal with challenges in face recognition in unconstrained videos. Deep Neural Networks (DNNs) have shown the ability to learn face representations that are robust to occlusions, image blur and large pose variations to achieve high recognition performance on semi-constrained still face recognition benchmarks [36, 40, 25, 39]. While face recognition in surveillance video and unconstrained still face images share similar challenges, video sequences from CCTV are generally of lower resolution and may contain noisy frames with poor quality and unfavorable viewing angles (See Figures 1 and 5 (d)). Such noisy frames will undermine the overall performance of video face recognition if we directly use recognition methods developed for still images.

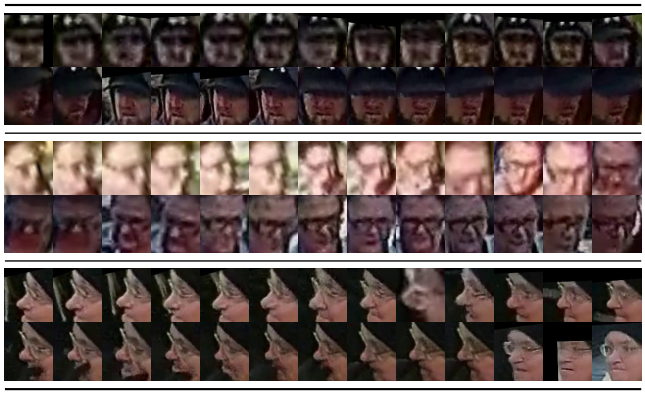


Figure 1: Example video frames of three subjects from IJB-S dataset [21]. Each subject has two rows of frames.

In this paper, we address unconstrained template-based[2] face recognition. Specifically, we consider the following five protocols in surveillance scenarios [21] (Figure 2): (i) surveillance-to-still, where the query is a surveillance video and each subject has a single frontal still image in the gallery; (ii) surveillance-to-booking, where each subject has a booking template (a set of still face images captured at enrollment) in the gallery; (iii) surveillance-to-surveillance; (iv) multi-view surveillance-to-booking, where the query is a collection of surveillance videos of one subject; (v) UAV surveillance-to-booking, where the query is a video captured by a small fixed-wing aerial vehicle.

State-of-the-art methods for face recognition in video represent a subject's face as an unordered set of vectors and the recognition is posed as estimating the similarity between face templates [1, 41, 5, 14, 50]. However, this is not computationally efficient as one needs to compare similarities on all feature vectors between two face templates. Thus, it

[1] Close to 200 million surveillance cameras have already been installed across China, which amounts to approximately 1 camera per 7 citizens. Approximately 40 million surveillance cameras were active in the United States in 2014, which amounts to approximately 1 camera per 8 citizens [42].

[2] A template is a set of images from the same person, first introduced in the IARPA Janus Benchmark [22]

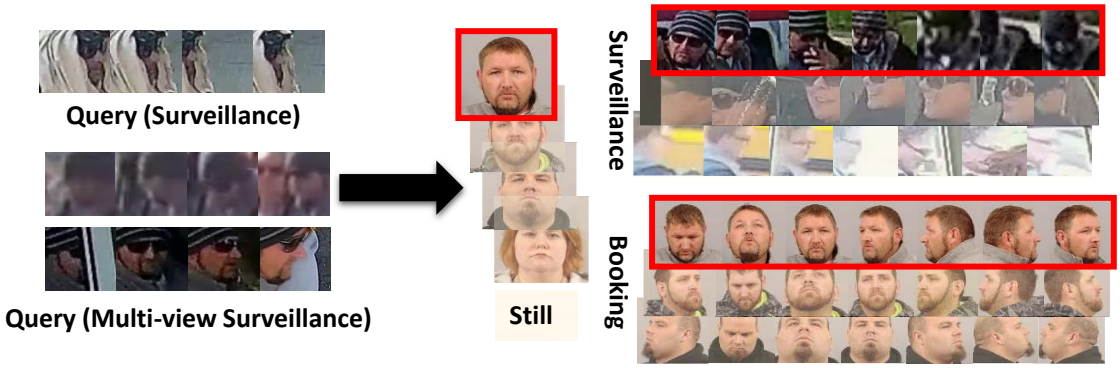
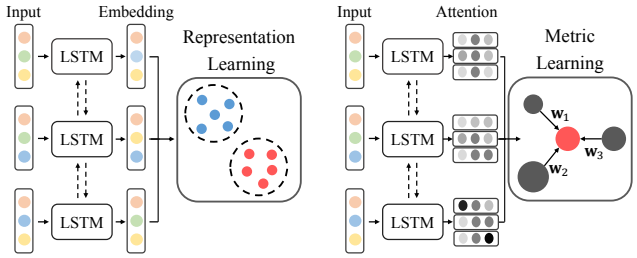


Figure 2: Surveillance video protocols in IJB-S dataset [21]. The queries are on the left and the galleries are on the right of the arrow. In each protocol, the query is either one surveillance video or multiple videos of the same person, which needs to be compared with every item in one of the three galleries types (still images, surveillance videos, and booking photos). The red boxes are used to highlight the ground truth identity. Since faces are difficult to detect in UAV videos, we only show examples for four of the five protocols.



(a) LSTM (b) MARN

Figure 3: Difference between naive LSTM and the proposed Multi-mode Aggregation Recurrent Network (MARN). LSTM learns new representations from scratch leading to overfitting. However, MARN utilizes pre-trained embeddings and integrates them using context information.

is preferable to aggregate feature vectors into a compact feature vector for each template [46, 44, 34, 38, 51, 26, 33, 27]. Most methods aggregate image sets based on the quality of each image or video frame. but ignore the contextual information and multi-modal attributes in video frames.

To exploit information in a set of face images or video frames, we propose a *Multi-mode Aggregation Recurrent Network (MARN)* for low quality video face recognition. RNN models capture information from sequential data via a memory mechanism to utilize the context information. However, directly using a RNN for learning video face representations could drop the discriminative features learned by embedding CNNs. In contrast, MARN learns to aggregate the embeddings of face images by leveraging the context-awareness of RNNs. Moreover, to disentangle features of different attributes or capture conditions, MARN introduces a multi-mode attention learning that maps weight vectors into multiple subspaces and aggregates features in each subspace separately.

Experimental results on the real-world low quality IJB-S dataset [21] and other template/video matching benchmarks show that the proposed MARN outperforms the face recognition performance of average pooling and other state-of-the-art aggregation methods. Specific contributions of the paper are listed below:

- A Multi-mode Aggregation Recurrent Network (MARN) that aggregates deep feature vectors based on contextual quality information in various modes (weights to aggregate image-based feature vectors instead of directly learning an aggregated representation), resulting in discriminative video face representations.
- The attention scores of one video frame provided by MARN present the relative discrimination power of different modes given the other frames in the video.
- State-of-the-art performance on a low quality surveillance benchmark IJB-S [21] and comparable results on two other face recognition benchmarks, YouTube Faces [43], and PaSC [4].

2. Related Work

2.1. Facial Analysis with RNN

Existing approaches for facial analysis of videos have utilized RNNs to account for the temporal dependencies in sequences of frames. For example, Gu *et al*. [12] proposed an end-to-end RNN-based approach for head pose estimation and facial landmark estimation in videos. As for recognition tasks, Ren *et al*. [35] attempted to address large out-of-plane pose invariant face recognition in image sequences by using a Cellular Simultaneous Recurrent Network (CSRN). Graves *et al*. [11] employed RNN that accepts a sequence of face features as input for facial expression recognition. RNN has also been widely used for face emotion recognition [49, 9, 8] in videos.

2.2. Video Face Recognition

State-of-the-art methods for video face recognition can primarily be put into three categories: *space*-model, *classifier*-model, and *aggregation*-model. Many traditional

space-models attempt to estimate a feature space where all the video frames can be embedded. Such a feature space can be represented as probabilistic distribution [37, 1], n^{th}-order statistics [28], affine hulls [5, 17, 47], SPD matrices [19], and manifolds [23, 14, 41]. *Classifier*-models [43, 30] learn face representations based on videos or image sets whereas *aggregation*-models strive to fuse the identity-relevant information in the face templates/videos to attain both efficiency and recognition accuracy. Best-Rowden *et al*. [3] showed that combining multiple sources of face media [3] boosts the recognition performance for identifying a person of interest. Most recent methods aim to aggregate a set of deep feature vectors into a single vector. Compared to simply averaging all vectors [7, 6], fusing features with the associated visual quality shows more promising results in recognizing faces in unconstrained videos. Ranjan *et al*. [32] utilized face detection scores as measures of face quality to rescale the face similarity scores. Yang *et al*. [46] and Liu *et al*. [27] proposed to use an additional network module to predict a quality score for each feature vector and aggregates the vectors weighted the assigned scores. Gong *et al*. [10] extended the aggregation model by considering component-wise quality prediction. Rao *et al*. [34] used LSTM to learn temporal features while use reinforcement learning to drop the features of low-quality images. [26] proposed a dependency-aware pooling by modeling the relationship of images within a set and using reinforcement learning for image quality prediction. None of these approaches have addressed redundancy in the video frames.

3. Motivation

Image quality of video frames obtained from deployed CCTV cameras is significantly lower than still images captured under constrained conditions. In addition, video frames may suffer severe motion blur and out-of-focus blur due to camera jitter and small oscillation in the scene. One way to address the large variations in face quality is to select key frames and eliminate poor quality images [31]. Hassner *et al*. [15] found that the recognition performance is undermined by removing low quality images.

To exploit information of an identity whose face images are sampled sequentially from a video, one simple idea is to linearly aggregate feature vectors extracted from images in a template by an adaptive weigthting scheme to generate a compact face representation for the template [46, 27, 45, 10]. We can formulate the feature aggregation in a probabilistic manner. The face space of noisy embeddings extracted by using a given facial representation model can be formulated as:

$$p(\mathbf{f}|\mathbf{I}^*, \mathbf{F}^*) = \int p(\mathbf{f}|\mathbf{i}, \mathbf{I}^*, \mathbf{F}^*) p(\mathbf{i}|\mathbf{I}^*, \mathbf{F}^*) d\mathbf{i}, \quad (1)$$

where $\mathbf{I}^* = \{\mathbf{i}_1, \mathbf{i}_2, \cdots, \mathbf{i}_M\}$ is the set of training images to learn the model parameters, $\mathbf{F}^* = \{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_M\}$ is the collection of noisy embeddings of the training data (feature vectors extracted from the image set $\mathbf{I}^*$), $p(\mathbf{f}|\mathbf{i}, \mathbf{I}^*, \mathbf{F}^*)$ is the uncertainty of embedding estimation given the training images, and $p(\mathbf{i}|\mathbf{I}^*, \mathbf{F}^*)$ is the probability density of face images in the underlying manifold of noiseless embeddings. In addition, we assume that there is a deterministic function that maps the face images of each identity to the corresponding noiseless embedding $\boldsymbol{\mu}$.

Let $T = \{\mathbf{i}_1, \mathbf{i}_2, \cdots, \mathbf{i}_N\}$ denote a template of one identity, where N is the number of images in the template. Since N can be large in case of videos, the noiseless embedding $\boldsymbol{\mu}$ can be approximated by the expectation $\hat{E}(\mathbf{F}^T)$:

$$\boldsymbol{\mu} \approx \hat{E}(\mathbf{F}^T) = \sum_{i=1}^{N} p(\mathbf{f}|\mathbf{I}^*, \mathbf{F}^*)\mathbf{f}_i, \quad (2)$$

where $\mathbf{F}^T$ is the collection of noisy embeddings in the template T. However, estimating the probabilistic density of face embeddings that can account for various sources of noise in face representations is challenging. An alternative solution is to estimate an adaptive scalar weight based on each feature vector (noisy embedding) and the approximated template embedding is the linear combination of the vectors based on the weights [46], [27]:

$$\mathbf{r}^T = \sum_{i=1}^{N} g(\mathbf{f}_i)\mathbf{f}_i, \quad (3)$$

where $\mathbf{r}^T$ is the template representation, and $g(\mathbf{f}_i)$ is the predicted weight for the feature vector of the i^{th} image in the template. Although this approach can reduce feature noise to some extent, the output weight is only inferred from the current feature vector. For videos captured by surveillance cameras, most of the face frames are corrupted and only a small proportion is of high quality. Given that frame qualities are seriously unbalanced, such a weighing scheme may still be affected by observational error. This motivated us to estimate quality weights based on context information by using all the images in a template.

In this paper, we propose an RNN guided feature aggregation network which predicts a quality weight for each deep feature vector based on the other vectors in the same template. Similar to [10], the proposed RNN-based model generates different weights for each component (dimension) of the deep feature vectors. Hence each component of the template representation is:

$$\mathbf{r}_j^T = \sum_{i=1}^{N} g(\mathbf{f}_{1j}, \mathbf{f}_{2j}, \cdots, \mathbf{f}_{Nj})\mathbf{f}_{ij}, \quad (4)$$

[3] Face media refers to a collection of sources of face information, for example, video tracks, multiple still images, 3D face models, verbal descriptions and face sketches.

where $\mathbf{r}_j^T$ is the j^{th} component of the template representation, and $g(\mathbf{f}_{1j}, \mathbf{f}_{2j}, \cdots, \mathbf{f}_{Nj})$ is the predicted weight for the j^{th} component of the feature vector of the i^{th} image in the template. By using context information, the overall influence of poor quality components is diminished; features with relatively large information can still benefit the final representation in spite of their lack of quantity.

Following [50], we consider multi-mode feature aggregation, where the weight vectors are mapped into a number of attention groups, instead of learning a single pooling mask. Each distinct group learns to employ features of a certain face attribute or condition that presents small intra-mode variance. In this way, features are disentangled into multiple modes increasing the effectiveness of aggregation.

4. Approach

4.1. Overall Framework

The overall framework is presented in Figure 4. A base CNN model is incorporated for extracting features from each face image and then MARN aggregates these features by considering information of each mode in the whole video sequence. We first train the base CNN on a large dataset of still face images, namely MS-Celeb-1M [13]. The learned model is used to extract the features from a video face dataset, UMDFaceVideo [22], which is further used to train the MARN to adaptively predict multiple weights for each deep feature. The feature vectors are aggregated in each individual mode and then concatenated into a single compact vector as the template representation.

4.2. Multi-mode Aggregation Recurrent Network (MARN)

Let $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_N\}$ be the set of CNN feature vectors representing face images in a template T, where each $\mathbf{f}_i$ is a D-dimensional vector and N is the number of face images in the template. A hidden state $\mathbf{h}_t$ of LSTM at time step t is computed based on the hidden state at time $t-1$, the input $\mathbf{f}_t$ and the cell state $\mathbf{C}_t$ at time t:

$$\mathbf{h}_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{f}_t] + b_o) \cdot tanh(\mathbf{C}_t), \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function and $\mathbf{W}_o$ and b_o are the parameters of the sigmoid gate. The attention vector of the k^{th} mode for the t^{th} feature vector in $\mathbf{F}$ is then inferred by the subsequent fully-connected layer: $\mathcal{H}(\mathbf{h}_t) = \mathbf{q}_t^k$, where the dimensionality of $\mathbf{q}_t^k$ is compressed to D/K and K is the number of aggregation modes. Correspondingly, the t^{th} feature vector is also compressed to D/K dimensions by $\mathcal{H}(\mathbf{f}_t) = \mathbf{f}_t^k$ for component-wise feature aggregation. A softmax operator normalizes all attention vectors of the same mode in the template along each component. Specifically, given a set of attention vectors of k^{th} mode $\{\mathbf{q}_1^k, \mathbf{q}_2^k, \cdots, \mathbf{q}_N^k\}$, the j^{th} component of the t^{th} vector is normalized by $w_{tj}^k = \frac{\exp(q_{tj}^k)}{\sum_{i=1}^N \exp(q_{ij}^k)}$. The template representation of k^{th} mode is the weighted mean vector of elements in $\mathbf{F}$:

$$\mathbf{r}^k = \sum_{i=1}^{N} \mathbf{f}_i^k \odot \mathbf{w}_i^k, \tag{6}$$

where $\odot$ denotes the element-wise multiplication, and the final template representation is the concatenation of all K $\mathbf{r}^k$ to obtain a D-dimensional feature vector for template T ($\mathbf{r}^T$).

4.3. Network Training

The architecture of MARN consists of a bi-directional LSTM networks with $2 * K$ layers and a fully-connected layer. The fully-connected layer is needed to project both CNN feature vectors and LSTM embeddings into the target dimension. To optimize the weight prediction, we adopt a template-based triplet loss. The triplet comprises one anchor template, one positive template of the same subject as the anchor, and one negative template of a different identity. All the templates are randomly selected to form a mini-batch and average hard triplet is utilized. Here, the hard triplet means the non-zero loss triplets [16]. The loss function is formulated as $\mathcal{L}_{triplet} = \frac{1}{M}\sum_{i=1}^{M}[\|\mathbf{r}_i^{T_a} - \mathbf{r}_i^{T_p}\|_2^2 - \|\mathbf{r}_i^{T_a} - \mathbf{r}_i^{T_n}\|_2^2 + \beta]_+$, where M is the number of hard triplets in a mini-batch, and $\{\mathbf{r}_i^{T_a}, \mathbf{r}_i^{T_p}, \mathbf{r}_i^{T_n}\}$ stands for the i^{th} triplet with anchor, positive, and negative template representations derived by Equation 6. $[x]_+ = \max(0, x)$, and β is the margin parameter.

5. Experiments

5.1. Datasets and Protocols

We train MARN on UMDFaceVideo dataset [2], and evaluate it on three other video face datasets (IJB-S [21], YTF [43], and PaSC [4]) without further fine-tuning.

UMDFaceVideo contains 3,735,476 annotated video frames extracted from a total of 22,075 videos of 3,107 subjects. The videos are collected from YouTube. The dataset is only used for training.

IJB-S is a surveillance video dataset collected by IARPA Janus program for unconstrained face recognition system evaluation. The dataset is composed of 350 surveillance videos with 30 hours of recording in total, 5,656 enrollment images, and 202 enrollment videos. The videos were captured under real-world environments to simulate law enforcement and security applications. The evaluations consist of five identification experiments as mentioned in section 1. We report the close-set Identification Rate (IR) and open-set performance in terms of TPIR @ FPIR [4]. Due to the poor quality of video frames, only 9 million out of 16

[4]True positive identification rate and false positive identification rate.

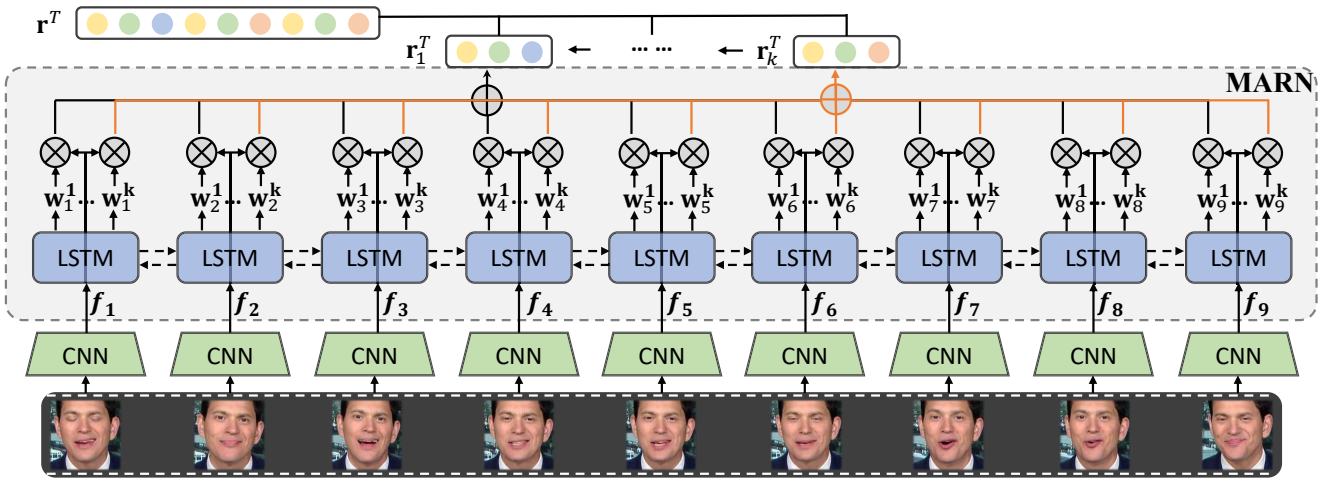


Figure 4: Overview of the proposed MARN. A sequence of face video frames are first input to a CNN model to extract deep face features. This is followed by bidirectional a LSTM model to predict multi-mode attention scores for each feature. The network finally outputs a single feature vector as the face representation of the set of frames for face recognition.

(a) YoutubeFaces [43] (b) PaSC [4] (c) UMDFaceVideo [22] (d) IJB-S [21]

Figure 5: Example images from four different video datasets. YoutubeFaces, PaSC and UMDFaceVideo contain only video frames. IJB-S includes both still images and real-world videos. The first column of IJB-S shows still images of three subjects, followed by video frames of the respective subjects in the next three columns.

million faces could be detected. Failure-to-enroll face images do not get utilized in template feature aggregation, and we use zero-vector as the template representation if no faces can be enrolled in the template.

YTF contains 3,425 videos of 1,595 different subjects. Unlike surveillance scenario of IJB-S, most videos in YTF are obtained from social media, where faces are more constrained and have higher quality. We report 1:1 face verification rate of the specific 5,000 video pairs.

PaSC contains 2,802 videos of 265 subjects. The dataset consists of two subsets of videos captured by control [5] and handheld cameras [6], respectively.

5.2. Implementation Details

Pre-processing: All the faces in the video are automatically detected by MTCNN [48]. Detected face regions are cropped from the original images and are resized into 112×96 after alignment by similarity transformation based on five facial landmarks [7] provided by MTCNN.

Training: Our base CNN model is a 64-layer residual network [25] trained on a clean version[8] of MS-Celeb-1M dataset [13] to learn a 512-dimensional face representation. The parameters of MARN are then trained on UMDFace-Video [2] with Adam optimizer, whose first momentum is set to 0.9 and the second momentum is 0.999. The margin of triplet loss is 3.0. During training, we define a template as the frames in the same video of one subject; the number of frames in each template is fixed. Each mini-batch incorporates 384 templates that are randomly sampled from 128 subjects with 32 images per template. The model is trained for 20 epochs. We remove the subjects which appear in the testing datasets; a small subset of the training set is held as validation set for tuning the hyper-parameters. We conduct all the experiments on a Nvidia Geforce GTX 1080 Ti GP; the average time of feature extraction is 1ms per image.

[5] A high quality Panasonic camera on a tripod
[6] Five hand held cameras with various resolutions
[7] Left eye, right eye, center of nose, left and right edge of mouth
[8] https://github.com/inlmouse/MS-Celeb-1M_WashList

5.3. Baseline

We design three baseline experiments to evaluate the proposed network, MARN.

- ***AvgPool*** uses average pooling of the base CNN features to generate the template representation.
- ***LSTM*** is a two-layer LSTM network to predict the template representation directly without attention based feature aggregation. The output of the last cell is used as the representation.
- ***QualityPool*** is a two-layer fully-connected network with ReLU as the activation function in between. Similar to previous work [46, 27, 10], the model also predicts quality weights and takes the weighted sum of all vectors as the template representation.

5.4. Ablation Study on Multi-mode Attention

In this section, we analyze the impact of using different number of modes for MARN. The recognition results on IJB-S dataset are reported in table 1. We observe that models with small or large number of modes (K) lead to poor identification performance. For small K values, the intra-mode variations are large. On the other hand, a large number of modes results in over-compressed observation vectors that may not provide sufficiently discriminative features. In the following experiments, we use the 4-mode MARN by default.

Table 1: Different number of modes on IJB-S.

Test Protocol	K^a	Closed-set (%)		Open-set (%)
		Rank-1	Rank-5	1.0 % FPIR
SV* to B†	1	58.52	65.21	31.25
	2	58.03	65.78	30.64
	4	**59.26**	**65.93**	**32.07**
	8	59.19	64.59	31.32
	16	57.86	64.16	31.88
SV* to SV*	1	21.96	33.79	**0.21**
	2	22.04	34.05	0.11
	4	**22.25**	**34.16**	0.19
	8	19.87	31.93	0.09
	16	19.22	32.57	0.06

[a] The number of aggregation modes
[*] Surveillance videos
[†] Booking images

5.5. Qualitative Analysis of IJB-S

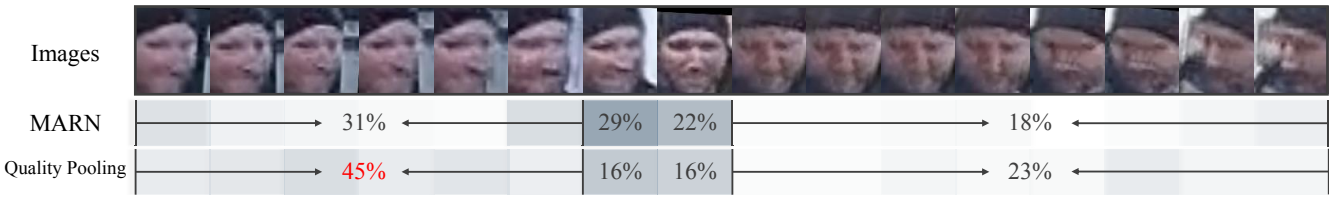
To evaluate the effect of context-aware pooling by MARN, we visualize the attention distribution on two example video sequences from IJB-S and PaSC. The two sequences are composed by randomly sampling 16 frames from the original video. Then the two models, *QualityPool* and MARN, are used to compute the attention for the images in the sequence. For visualization purpose, the attention vector (for different components) of each image is averaged into one scalar. Two example results are shown in Figure 6. While *QualityPool* can effectively predict the quality of the given image, but without context information, it is easily distracted when the number of identical (with little change in the frames) video frames is large. The first 6 frames in the IJB-S video are nearly identical in content, but overall they have a larger weight than the two higher quality faces. Similarly, in the PaSC case, the last 5 frames almost contain the same information, but overall they receive 41% attention. However, MARN is robust against redundancy.

5.6. Quantitative Analysis on IJB-S

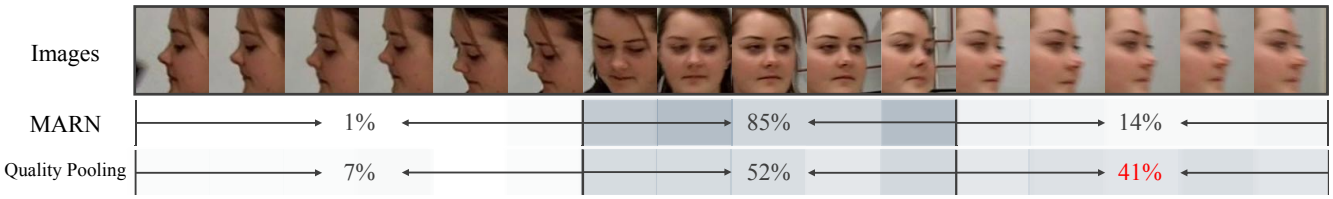
Table 2: Comparisons of MARN with baselines on IJB-S.

Test Name	Method	Closed-set (%)		Open-set (%)
		Rank-1	Rank-5	1.0 % FPIR
SV to still	C-FAN [10]	50.82	61.16	16.44
	AvgPool	50.80	59.60	11.60
	LSTM	2.90	17.46	0.12
	QualityPool	51.61	62.78	17.33
	MARN	**58.14**	**64.11**	**21.47**
SV to B	C-FAN [10]	53.04	62.67	27.40
	AvgPool	50.82	59.73	19.19
	LSTM	4.49	16.40	1.31
	QualityPool	52.66	62.95	25.60
	MARN	**59.26**	**65.93**	**32.07**
Multi-view SV to B	C-FAN [10]	96.04	99.50	70.79
	AvgPool	96.53	98.51	66.33
	LSTM	4.95	19.80	3.21
	QualityPool	97.03	99.50	75.62
	MARN	**98.87**	**99.50**	**76.89**
SV to SV	C-FAN [10]	10.05	17.55	0.11
	AvgPool	7.71	14.34	0.08
	LSTM	11.13	21.07	0.08
	QualityPool	5.69	9.43	0.09
	MARN	**22.25**	**34.16**	**0.19**
UAV SV to B	C-FAN [10]	7.59	**12.66**	0.00
	AvgPool	2.53	6.33	0.00
	LSTM	1.30	2.66	0.00
	QualityPool	7.59	10.85	0.00
	MARN	**7.63**	12.28	**3.13**

Table 2 reports identification results on the five protocols of the IJB-S dataset. The proposed approach achieves better performance than the baselines in nearly all cases. In particular, the proposed approach outperforms C-FAN [10] by 11.91% and 15.24% on the closed-set surveillance-to-surveillance protocol at rank 1 and rank 5, respectively. Moreover, with the proposed technique, the open-set protocol results are also improved by 3.62%, 3.85%, and 5.45% on surveillance-still, surveillance-to-booking, and multi-view surveillance-to-booking at 1% FPIR, respectively. It is worth noting that MARN is trained on video frames data with temporal information. However, the booking images in IJB-S are still face images with high quality, on which the quality prediction is not as useful as surveillance



(a) Example from IJB-S



(b) Example from PaSC

Figure 6: Attention distribution of one-mode MARN and quality-aware pooling on two video sequences from IJB-S and PaSC. The summation of the weight percentage is given below each clustering of the frames. The darker the blue box is, the higher the attention weight of the corresponding frames. Although "QualityPool" is able to assign larger weights to higher quality images, it can be distracted by redundant low-quality images if there is a sufficiently large fraction of them (red numbers). MARN stays focused on high quality frames via the context information. The attention is computed by averaging across all components of the weight vector.

Table 3: Verification Accuracy on YTF

Method	Accuracy (%)	Method	Accuracy (%)
EigenPEP [24]	84.8 ± 1.4	DeepFace [40]	91.4 ± 1.1
DeepID2+ [39]	93.2 ± 0.2	C-FAN [10]	$\mathbf{96.50 \pm 0.90}$
FaceNet [36]	95.52 ± 0.06	DAN [33]	94.28 ± 0.69
NAN [46]	95.72 ± 0.64	QAN [27]	96.17 ± 0.09
AvgPool	96.24 ± 0.96	*LSTM*	60.00 ± 2.81
QualityPool	96.38 ± 0.95	**MARN**	96.44 ± 0.99

frames. Therefore, for booking images, we use average pooling instead of quality aggregation. In comparison with the three baseline methods, MARN achieves higher identification rates than *LSTM* on all five protocols. Obviously, the *LSTM* over fits to the video frames in UMDFaceVideo and is not able to generalize to the booking images. *QualityPool* achieves similar performance as C-FAN, since both of the approaches use component-wise attention scheme. Both *QualityPool* and the proposed MARN outperform average pooling of the base CNN features.

5.7. Performance Comparison on YTF and PaSC

Table 3 reports the face verification performance of the proposed method and other state-of-the-art methods on YTF dataset. MARN outperforms all of our three baselines. The performance of MARN is slightly higher than previous approaches such as NAN [46] and C-FAN [10]. Since YouTube Face videos are not captured by typical surveillance cameras, instead, most of them are recorded by professional photographers. As a result, they are free from very low quality frames. For this reason, the proposed context-aware attention network does not offer obvious advantages over some of the state-of-the-art approaches.

Table 4: Comparisons of the verification rate (%) on PaSC at a false accept rate (FAR) of 0.01.

Method	Control	Handheld
DeepO2P [29]	68.76	60.14
SPDNet [18]	80.12	72.83
GrNet [20]	80.52	72.76
Rao *et al.* [34]	95.67	93.78
TBE-CNN [7]	95.83	94.80
TBE-CNN + BN [7]	**97.80**	**96.12**
AvgPool	91.27	74.30
LSTM	3.07	1.28
QualityPool	96.48	92.39
MARN	96.67	95.13

Table 4 reports the verification results on PaSC dataset. In comparison with YTF, PaSC is more challenging since the faces in the dataset have full pose variations. By comparing the proposed MARN with the three baseline models, we can observe that combining context information with attention aggregation is capable of improving the discriminative power of video face representations. We can also observe that the proposed approach achieves comparable performance to other state-of-the-art methods.

6. Conclusions

To address face recognition in low quality unconstrained surveillance videos, we propose a Multi-mode Aggregation Recurrent Network (MARN) that adaptively predicts context-aware quality weight vectors for each deep feature vector extracted by CNN face model. Each face in a video

is represented as a compact deep feature vector aggregated by MARN under the weighted attention scheme. Experimental results on three video face datasets, i.e., IJB-S, YTF, and PaSC show that the attention values provided by the proposed MARN enables utilizing discriminative features while discarding the noisy features by leveraging the context information in the video learned by LSTM. Our method shows advantages on video face benchmarks, especially low quality videos in IJB-S benchmark.

References

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*, 2005. 1, 3

[2] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa. The do's and don'ts for cnn-based face verification. In *ICCV Workshops*, 2017. 4, 5

[3] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Trans. on Information Forensics and Security*, 2014. 3

[4] 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, et al. The challenge of face recognition from digital point-and-shoot cameras. In *BTAS*, 2013. 2, 4, 5

[5] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, 2010. 1, 3

[6] J.-C. Chen, R. Ranjan, S. Sankaranarayanan, A. Kumar, C.-H. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa. Unconstrained still/video-based face verification with deep convolutional neural networks. *IJCV*, 2018. 3

[7] C. Ding and D. Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Trans. on PAMI*, 2018. 3, 7

[8] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *ACM MM*, 2015. 2

[9] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450, 2016. 2

[10] S. Gong, Y. Shi, and A. K. Jain. Video face recognition: Component-wise feature aggregation network (c-fan). In *IEEE ICB*, 2019. 3, 6, 7

[11] A. Graves, C. Mayer, M. Wimmer, J. Schmidhuber, and B. Radig. Facial expression recognition with recurrent neural networks. In *International Workshop on Cognition for Technical Systems*, 2008. 2

[12] J. Gu, X. Yang, S. De Mello, and J. Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *CVPR*, 2017. 2

[13] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large scale face recognition. In *ECCV*, 2016. 4, 5

[14] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *CVPR*, 2011. 1, 3

[15] T. Hassner, I. Masi, J. Kim, J. Choi, S. Harel, P. Natarajan, and G. Medioni. Pooling faces: template based face recognition with pooled face images. In *CVPR Workshops*, 2016. 3

[16] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017. 4

[17] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, 2011. 3

[18] Z. Huang and L. Van Gool. A riemannian network for spd matrix learning. In *AAAI*, 2017. 7

[19] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *ICML*, 2015. 3

[20] Z. Huang, J. Wu, and L. Van Gool. Building deep networks on grassmann manifolds. In *AAAI*, 2018. 7

[21] N. D. Kalka, B. Maze, J. A. Duncan, K. J. OConnor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain. IJB-S : IARPA Janus Surveillance Video Benchmark . In *BTAS*, 2018. 1, 2, 4, 5

[22] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *CVPR*, 2015. 1, 4, 5

[23] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, 2003. 3

[24] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-pep for video face recognition. In *ACCV*, 2014. 7

[25] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 1, 5

[26] X. Liu, B. Vijaya Kumar, C. Yang, Q. Tang, and J. You. Dependency-aware attention control for unconstrained face recognition with image sets. In *ECCV*, 2018. 2, 3

[27] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017. 2, 3, 6, 7

[28] J. Lu, G. Wang, and P. Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, 2013. 3

[29] E. Mohagheghian. *An application of evolutionary algorithms for WAG optimisation in the Norne Field*. PhD thesis, Memorial University of Newfoundland, 2016. 7

[30] M. Parchami, S. Bashbaghi, E. Granger, and S. Sayed. Using deep autoencoders to learn robust domain-invariant representations for still-to-video face recognition. In *IEEE AVSS*, 2017. 3

[31] X. Qi, C. Liu, and S. Schuckers. Cnn based key frame extraction for face in video recognition. In *International Conference on Identity, Security, and Behavior Analysis (ISBA)*, 2018. 3

[32] R. Ranjan, A. Bansal, H. Xu, S. Sankaranarayanan, J.-C. Chen, C. D. Castillo, and R. Chellappa. Crystal loss

and quality pooling for unconstrained face verification and recognition. *arXiv:1804.01159*, 2018. 3

[33] Y. Rao, J. Lin, J. Lu, and J. Zhou. Learning discriminative aggregation network for video-based face recognition. In *ICCV*, 2017. 2, 7

[34] Y. Rao, J. Lu, and J. Zhou. Attention-aware deep reinforcement learning for video face recognition. In *ICCV*, 2017. 2, 3, 7

[35] Y. Ren, K. Anderson, K. Iftekharuddin, P. Kim, and E. White. Pose invariant face recognition using cellular simultaneous recurrent networks. In *UCNN*, 2009. 2

[36] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1, 7

[37] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *ECCV*, 2002. 3

[38] K. Sohn, S. Liu, G. Zhong, X. Yu, M.-H. Yang, and M. Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. In *ICCV*, 2017. 2

[39] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, 2015. 1, 7

[40] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1, 7

[41] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*, 2008. 1, 3

[42] Wikipedia. Mass surveillance in china. https://en.wikipedia.org/wiki/Mass_surveillance_in_China, 2019-03-17. 1

[43] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011. 2, 3, 4, 5

[44] W. Xie, L. Shen, and A. Zisserman. Comparator networks. In *ECCV*, 2018. 2

[45] W. Xie and A. Zisserman. Multicolumn networks for face recognition. *arXiv:1807.09192*, 2018. 3

[46] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *CVPR*, 2017. 2, 3, 6, 7

[47] M. Yang, P. Zhu, L. Van Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. In *IEEE FG*, 2013. 3

[48] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016. 5

[49] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li. Spatial-temporal recurrent neural network for emotion recognition. *IEEE Trans. on cybernetics*, 2018. 2

[50] J. Zhao, J. Li, X. Tu, F. Zhao, Y. Xin, J. Xing, H. Liu, S. Yan, and J. Feng. Multi-prototype networks for unconstrained set-based face recognition. *arXiv:1902.04755*, 2019. 1, 4

[51] Y. Zhong, R. Arandjelović, and A. Zisserman. Ghostvlad for set-based face recognition. *arXiv:1810.09951*, 2018. 2