Recurrent Embedding Aggregation Network for Video Face Recognition

Sixue Gong, Yichun Shi, Anil K. Jain

Michigan State University, East Lansing, MI

`gongsixu@msu.edu, shiyichu@msu.edu, jain@cse.msu.edu`

Abstract

Recurrent networks have been successful in analyzing temporal data and have been widely used for video analysis. However, for video face recognition, where the base CNNs trained on large-scale data already provide discriminative features, using Long Short-Term Memory (LSTM), a popular recurrent network, for feature learning could lead to overfitting and degrade the performance instead. We propose a Recurrent Embedding Aggregation Network (REAN) for set to set face recognition. Compared with LSTM, REAN is robust against overfitting because it only learns how to aggregate the pre-trained embeddings rather than learning representations from scratch. Compared with quality-aware aggregation methods, REAN can take advantage of the context information to circumvent the noise introduced by redundant video frames. Empirical results on three public domain video face recognition datasets, IJB-S, YTF, and PaSC show that the proposed REAN significantly outperforms naive CNN-LSTM structure and quality-aware aggregation methods.

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 video. Face recognition approaches for high quality still images (controlled capture and cooperative subjects) are not able to deal with challenges in face recognition in unconstrained videos. On the other hand, a video sequence contains more information about the subject in terms of temporal context (varying pose, expression, and motion) than still images. Thus, one of the key challenges in video face recognition is how to effectively combine facial information available across multiple video frames to improve face

---

[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 [43].
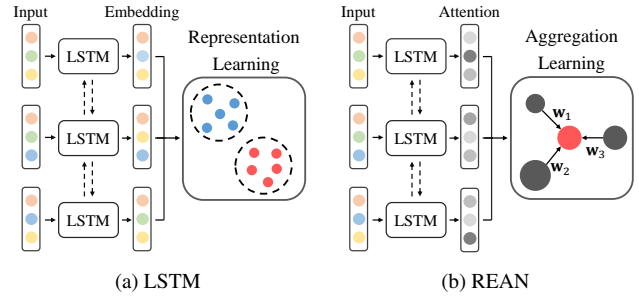

(a) LSTM (b) REAN

Figure 1: Difference between naive LSTM and the proposed Recurrent Embedding Aggregation Network (REAN). LSTM needs to learn new representations from scratch and hence could easily overfit. In comparison, REAN utilizes pre-trained embeddings and integrates them using context information.

recognition accuracy.

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 [37, 41, 26, 40]. While face recognition in surveillance video and unconstrained still face images share similar challenges, video sequences from CCTV are generally of low resolution and may contain noisy frames with poor quality and unfavorable viewing angles (See Figures 2 (d) and 3). Such noisy frames will undermine the overall performance of video face recognition if we directly use recognition methods developed for still images.

State-of-the-art methods for face recognition in video represent a subject's face as an unordered set of vector and the recognition is posed as estimating the similarity between face templates [1, 5, 15, 42, 51]. However, this is not computationally efficient as one needs to compare similarities on all feature vectors between two face templates. Thus, it is preferable to aggregate feature vectors into a compact feature vector for each template [47, 52, 27, 45, 34, 39, 35, 28]. Most methods aggregate image sets based on the quality of each image or video frame but ignores the contextual information.

In this paper, we propose a *Recurrent Embedding Aggregation Network (REAN)* for video face recognition. LSTM

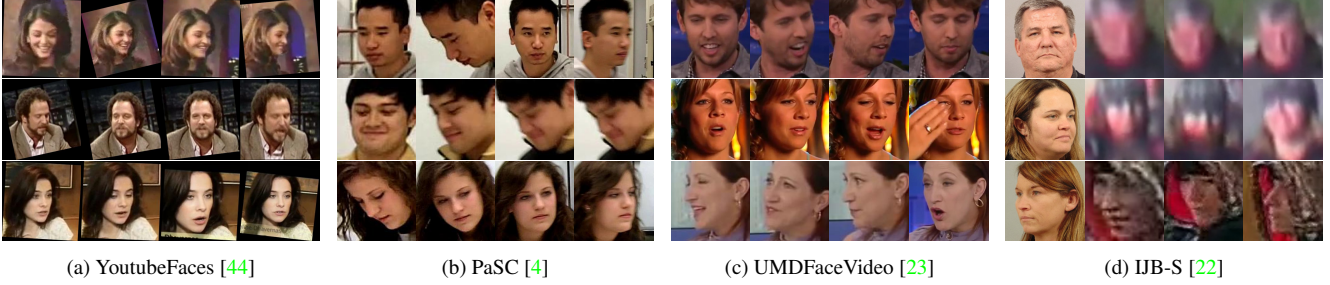| (a) YoutubeFaces [44] | (b) PaSC [4] | (c) UMDFaceVideo [23] | (d) IJB-S [22] |

Figure 2: Example images from different video datasets. YoutubeFace, PaSC and UMDFaceVideo contain only video frames. IJB-S includes both still images and videos. The first column of IJB-S shows still images, followed by video frames of the respective subjects in the next three columns.

models have been shown to robustly grasp information from sequential data by introducing a memory mechanism to utilize the context information. In spite of this, directly using a LSTM for learning video face representations could lead to degraded performance by dropping the discriminative features learned by embedding CNNs. Instead, the proposed REAN learns to aggregate the embeddings of face images by leveraging the context-awareness of LSTMs. Instead of a feature vector for matching, the LSTM in our case outputs an attention vector, which can be used for integration. Experimental results on IJB-S dataset [22] and other template/video matching benchmarks show that the proposed REAN significantly improves the face recognition performance compared with average pooling and other state-of-the-art aggregation methods. Specific contributions of the paper are listed below:

- A Recurrent Embedding Aggregation Network (REAN) that aggregates deep feature vectors based on quality and context information, resulting in better representation for video face recognition.

- The attention scores of one video frame provided by REAN present the relative discrimination power given the other frames in the video.

- REAN achieves state-of-the-art performance on a surveillance benchmark IJB-S [22] and two other face recognition benchmarks, YouTube Faces [44], and PaSC [4].

2. Related Work

2.1. Facial Analysis with RNN

Many 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. [13] proposed an end-to-end RNN-based approach to head pose estimation and facial landmark estimation in videos. As for recognition tasks, Ren et al. [36] 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. [12] employed RNN that accepts a sequence of Candide face features as input for facial expression recognition. RNN has also been widely used for face emotion recognition [50, 10, 9] from videos.

2.2. Video Face Recognition

State-of-the-art methods for video face recognition can primarily be summarized 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 feature space can be represented as probabilistic distribution [38, 1], n-order statistics [29], affine hulls [18, 5, 48], SPD matrices [20], and manifolds [24, 15, 42]. *classifier*-models [44, 31] are proposed to learn face representations based on videos or image sets; *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 [2] boosts the recognition performance for identifying a person of interest. With the ubiquity of deep learning algorithms for face recognition, most recent methods aim to aggregate a set of deep feature vectors into a single vector. Compared to simply averaging all vectors [8, 6], fusing features with the associated visual quality shows more promising results in recognizing faces in unconstrained videos. Ranjan et al. [33] utilized face detection scores as measures of face quality to rescale the face similarity scores. Yang et al. [47] and Liu et al. [28] proposed to use an additional network module that predicts a quality score for each feature vector and aggregates the vectors with the assigned scores. Gong et al. [11] extended the aggregation model by considering component-wise quality prediction Rao et al. [35] used LSTM to learn temporal features while use reinforcement learning to drop the features of low-quality images. [27] proposed a dependency-aware

---

[2]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.

pooling by modeling the relationship of images within a set and using reinforcement learning for image quality prediction. However, none of these approaches have address the redundancy issue in the video frames.

3. Motivation

3.1. Context Information

A face video provides a sequence of frames that can be used to perform recognition. This allows us to exploit contextual and temporal information of an identity whose face images are sampled sequentially from a video. For example, a variety of face poses and facial expressions can be present in the trajectory of the face movements in a video. On the other hand, the image quality of video frames obtained from typical CCTV cameras tend to be lower than still images captured under constrained conditions (e.g. passport photos). 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 [32]. Despite its efficiency, Hassner *et al.* [16] found that the overall recognition performance is undermined by simply removing low quality images.

Popular formulations of facial information integration in videos or image sets follow the idea of linearly aggregating feature vectors extracted from images in a template by an adaptive weighting scheme to generate a compact face representation for the template [47, 28, 46, 11]. A probabilistic model for the face space of noisy embeddings $\mathbf{y}$ extracted from a given facial representation model can be formulated as:

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

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

Suppose it is possible to capture a sufficient number of image samples of one identity to form a template $T = \{\mathbf{i}_1, \mathbf{i}_2, \cdots, \mathbf{i}_N\}$, 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 $\hat{E}(\mathbf{Y}^T)$:

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

where $\mathbf{Y}^T$ is the collection of noisy embeddings in the template T. However, estimating the probabilistic density of face embeddings while justifying various sources of noise in face representations is a very challenging task. An alternative solution for approximating the template representation is to estimate an adaptive scalar weight based on each feature vector (noisy embedding) and the approximated embedding of the template is the linear combination of the vectors based on the weights [47], [28]:

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

where $\mathbf{r}^T$ is the template representation, and $f(\mathbf{y}_i)$ is the predicted weight for the feature vector of the ith image in the template. Although this kind of approach can reduce feature noise to some extent, the output weight is only inferred from the current feature vector. Considering a situation in which a video is captured under unconstrained conditions, e.g., surveillance cameras, most of the face frames are corrupted and only a small proportion has relatively high quality. Since the observed frame qualities are seriously unbalanced, such a weight estimation scheme may still be affected by observational error. In data assimilation for numerical weather prediction (NWP), it is important to build an observation bias correction scheme based on intricate (spatial and temporal) knowledge of the systemic errors [7]. This inspires us to estimate quality weights based on context information by using all the images in a template to further diminish the impact of observational error in video clips. It has been shown that RNN is capable of capturing contextual cues by traversing the sequential slices along different directions [12].

In this paper, we propose an RNN guided feature aggregation network which predicts a quality weight for each deep feature vector based on vectors in the same template. Similar to [11], the proposed RNN-based model also generates different weights for each component (dimension) of the deep feature vectors. Hence each component of the template representation is:

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

where $\mathbf{r}_j^T$ is the jth component of the template representation, and $f(\mathbf{y}_{1j}, \mathbf{y}_{2j}, \cdots, \mathbf{y}_{Nj})$ is the predicted weight for the jth component of the feature vector of the ith image in the template. By using the context information, the weight value for each feature depends on other features in the template. Furthermore, the overall influence of poor quality components is diminished rather than enhanced by their large proportion in the template; frames with good quality can still dominate the final representation in spite of their lack of quantity.
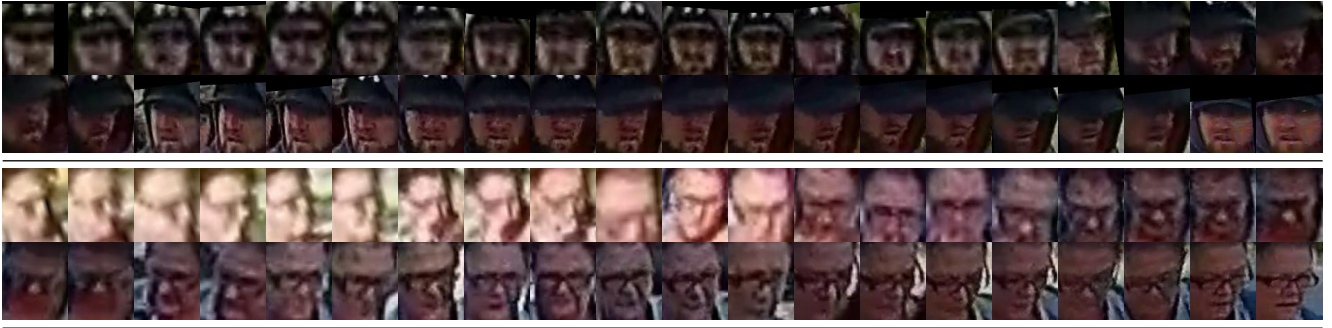
Figure 3: Example video frames from IJB-S dataset. We only show videos of two subjects due to the space limit. The first two rows show the frames from one subject and the last rows are frames of another subject.

Table 1: Face Identification Accuracy (both for closed-set and open-set scenarios) of feature aggregation with and without context information on five protocols of IJB-S dataset.

Test Protocol	Method	Closed-set (%)		Open-set (%)
		Rank-1	Rank-5	1 % FPIR
SV* to still	No-Context	49.57	59.58	16.48
	Context	**52.14**	**61.46**	**20.59**
SV* to B†	No-Context	51.14	61.43	26.64
	Context	**53.81**	**63.23**	**30.39**
Multi-view SV* to B†	No-Context	94.55	99.01	61.39
	Context	**97.52**	**99.01**	**74.26**
SV* to SV*	No-Context	**8.90**	**16.61**	0.06
	Context	8.85	16.55	**0.09**
UAV SV* to B†	No-Context	5.06	11.39	0.00
	Context	**6.33**	**13.92**	0.00

* Surveillance videos
† Booking images (the full set of images, frontal and profile, taken of a subject at enrollment time)

3.2. Illustration of Feature Aggregation

In this section, we provide an example to demonstrate the efficacy of feature aggregation by adopting context information in a template. Using a pre-trained CNN embedding [3] for extracting facial features, we train a multilayer perceptron (MLP) with two fully connected layers to predict a quality indicator for each feature vector in a similar manner as [47]. Then two types of aggregation strategies are considered:

- *No-Context*: The quality scores are normalized with softmax function and the fused representation is obtained by aggregating all the feature vectors with the normalized scores.

- *Context*: The images are separated into two groups using k-means on the quality scores. The fused representation is obtained by only aggregating the group with higher quality.

Table 1 reports the face recognition results of the above two aggregation strategies on IJB-S [22] under five different protocols (Section 5). It can be observed that the context information improves the face recognition performance on all five protocols, except the closed-set surveillance to surveillance protocol. For *No-Context*, since most of the frames in IJB-S videos contain redundant and low quality faces (See Figure 3), noisy embeddings in the template could still have a large impact because of their large proportion even when they are weighted by the quality scores. On the other hand, *Context* is able to further improve the performance by simply dropping the potential redundant images (images predicted with lower quality). Note that for the surveillance to surveillance protocol, both probe set and gallery set are composed of noisy frames. Therefore, by removing low quality frames has little impact on the performance.

4. Approach

4.1. Overall Framework

The overall framework of the proposed context-aware attention network is presented in Figure 4. A base CNN model is incorporated for extracting features from each face image and then the proposed Recurrent Embedding Aggregation Network (REAN)) is used to aggregate these features by considering the information in the whole video. The two models can be trained simultaneously in an end-to-end strategy, or separately. In this work, we first train the base CNN on subjects in a large dataset of still face images (MS-Celeb-1M [14]). Then it is used to extract the features from a video face dataset, UMDFaceVideo [23], which is further used to train the REAN to adaptively predict quality weights for each deep feature vector. The features in a template are then aggregated with context-aware quality weights into a single compact feature vector.

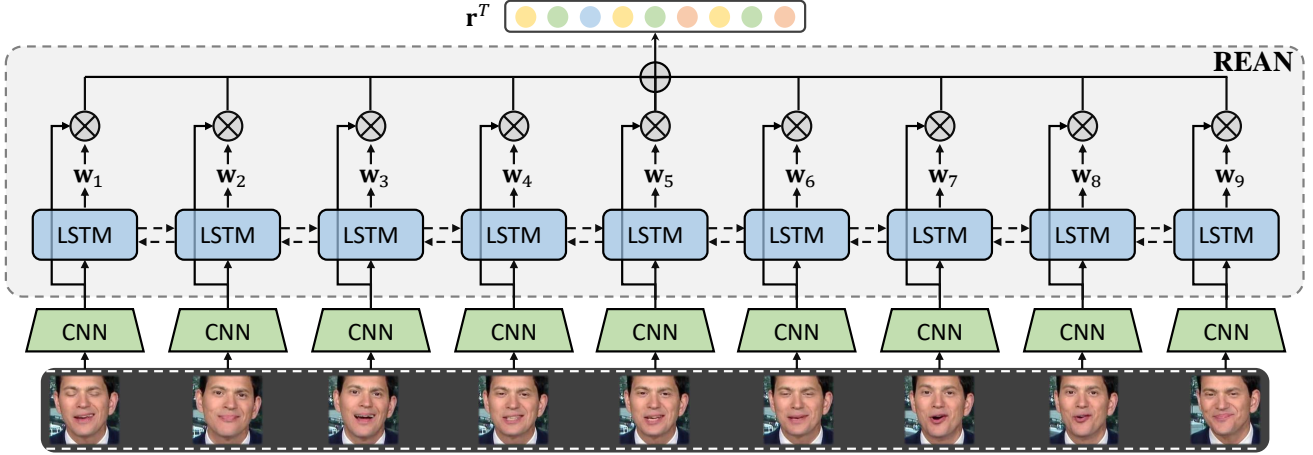[3] A 64-layer CNN [26] trained on MS-Celeb-1M [14] dataset.

Figure 4: Overview of the proposed REAN. A set of face video frames are first accepted by a CNN model to extract deep face features. Then a bidirectional LSTM model is fed by these deep feature vectors to predict attention score for each feature. The network finally outputs a single feature vector as the face representation of the set for the recognition task.

4.2. Context-aware Aggregation Module

Let $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_N\}$ be the CNN feature vectors of 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:

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

where $\sigma(\cdot)$ is the sigmoid function, $\mathbf{W}_o$ and b_o are the parameters of the sigmoid gate. The quality vector for the tth feature vector in $\mathbf{F}$ is then inferred by the subsequent fully-connected layer: $\mathcal{H}(\mathbf{h}_t) = \mathbf{q}_t$, where $\mathbf{q}_t$ has the same dimensionality D as $\mathbf{f}_t$. For the feature aggregation, we first employ a softmax operator to normalize all quality vectors in the template along each component. Specifically, given a set of quality vectors $\{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_N\}$, the jth component of the tth vector is normalized by:

$$w_{tj} = \frac{\exp(q_{tj})}{\sum_{i=1}^{N} \exp(q_{ij})} \qquad (6)$$

The final template representation is the weighted mean vector of elements in $\mathbf{F}$:

$$\mathbf{r}^T = \sum_{i=1}^{N} \mathbf{f}_i \odot \mathbf{w}_i, \qquad (7)$$

where $\odot$ denotes the element-wise multiplication, and $\mathbf{r}^T$ is a D-dimensional feature vector for template T.

4.3. Network Training

The architecture of REAN consists of a two-layer bidirectional LSTM network and a fully-connected layer. The fully-connected layer is needed to project the LSTM embeddings into the same dimension as the input feature vector. Note that the output of this module is the quality weights for each feature vector and the original CNN features is not modified. Finally, we aggregate the CNN features in the same template by leveraging the predicted weights for each component of the vector and use the weighted mean vector as the template representation. 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 [17]. The loss function is formulated as:

$$\mathcal{L}_{triplet} = \frac{1}{M} \sum_{i=1}^{M} [\|\mathbf{r}_i^{T_a} - \mathbf{r}_i^{T_p}\|_2^2 - \|\mathbf{r}_i^{T_a} - \mathbf{r}_i^{T_n}\|_2^2 + \beta]_+,$$

$$(8)$$

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 ith triplet with anchor, positive, and negative template representations derived by Equation 7. $[x]_+ = \max(0, x)$, and β is the margin parameter.

5. Experiments

5.1. Datasets and Protocols

We train REAN on UMDFaceVideo dataset [2], and evaluate it on three other video face datasets, the IARPA Janus Benchmark - Surveillance (IJB-S) [22], the YouTube Face dataset (YTF) [44], and the Point-and-Shoot Challenge dataset (PaSC) [4].

UMDFaceVideo contains 3,735,476 annotated video frames extracted from a total of 22,075 videos for 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 for face recognition system evaluation. The dataset is composed of 350 surveillance videos with 30 hours in total, 5,656 enrollment images, and 202 enrollment videos. The videos were captured under real-world environments, and can be used to simulate law enforcement and security applications. The dataset also contains several surveillance relevant protocols, including closed-set and open-set 1:N identification evaluations. The evaluations consist of five experiments: i) surveillance-to-still [4], ii) surveillance-to-booking [5], iii) surveillance-to-surveillance, iv) multi-view surveillance-to-booking, and v) UAV [6] surveillance-to-booking. IJB-S also includes a face detection protocol. Since our work mainly focuses on recognition and not detection, we only follow the five identification protocols and report the standard Identification Rate (IR) and open-set performance in terms of TPIR @ FPIR [7]. Due to the poor quality of video frames in IJB-S, only 9 million out of 16 million faces could be detected. In our experiments, 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 is a video face dataset that contains 3,425 videos of 1,595 different subjects. The average number of frames in YTF videos is 181. Unlike surveillance scenario of IJB-S, most videos in YTF are from video capture or posted on social media, where faces are more constrained and have higher quality. We report the 1:1 face verification rate of the given 5,000 video pairs in our experiments without fine-tuning.

PaSC is a video face dataset that contains 2,802 videos of 265 subjects, varying in viewpoints, sensor types, and distance to the camera, etc. The dataset consists of two subsets of videos captured by control and handheld cameras, respectively. Similar to IJB-S, no training set is provided by PaSC, so we directly evaluate the proposed approach on PaSC without further pre-processing or fine-tuning.

5.2. Implementation Details

Pre-processing: All the face images in our experiments are automatically detected by a facial landmark detection algorithm, MTCNN [49]. 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 [8].

Training: Our base CNN model is a 64-layer residual network [26] trained on a clean version[9] of MS-Celeb-1M dataset [14] to learn a 512-dimensional face representation of still images. The parameters of REAN are then trained on UMDFaceVideo [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 set to 3.0. During training, we define one template as the frames in the same video of one subject and we fix the number of frames in one template. In particular, each mini-batch incorporates 384 templates that are randomly sampled from 128 subjects, 32 images per template. The model is trained for 20 epochs in total. We remove the subjects appear in the testing datasets and a small set of the training set is separated as validation set for tuning the hyper-parameters. We conduct all the experiments on a Nvidia Geforce GTX 1080 Ti GPU and the average speed of feature extraction is 1ms per image. No further fine-tuning is employed.

5.3. Baseline

We design three baseline experiments to compare with the proposed REAN.

- **AvgPool** simply applies the average pooling to 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 [47, 28, 11], the model also predicts quality weights for each feature vector and takes the weighted summation of all vectors as the template representation.

5.4. Qualitative Analysis of IJB-S

To evaluate the effect of context-aware pooling by REAN, we visualize the attention distribution on two example video sequences from IJB-S and PaSC. The two sequences are composed by randomly sampling 16 images from the original video. Then the two models, *QualityPool* and REAN, 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 5. *QualityPool* can effectively predict the quality of the given image, but without context information, it is

---

[4]"Still" stands for single frontal still images.

[5]The "booking" template comprises the full set of images captured of a single subject at enrollment time.

[6]UAV is a small fixed-wing unmanned aerial vehicle that was flown to collect images and videos.

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

[8]Left eye, right eye, center of nose, left edge of mouth and right edge of mouth

[9]https://github.com/inlmouse/MS-Celeb-1M_WashList
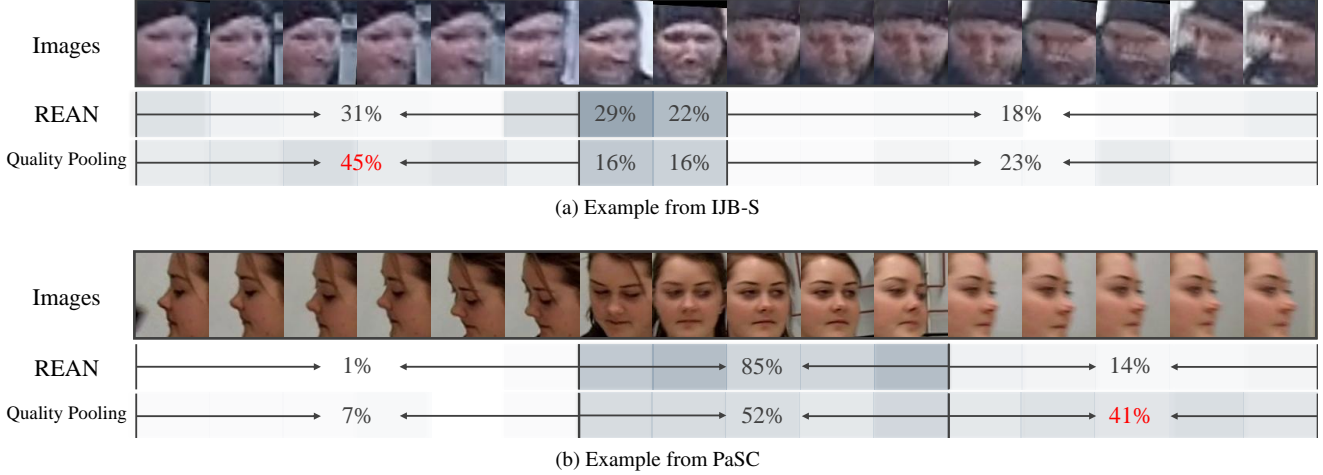
(a) Example from IJB-S

(b) Example from PaSC

Figure 5: Attention distribution of REAN 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 frame above. Although "Quality Pooling" is able to assign larger weights to higher quality images, it can be distracted by redundant low-quality images if there is a large number of them (red numbers). In comparison, REAN is able to keep focused on high quality frames by leveraging the context information. The attention here is computed by averaging across all components of the weight vector.

vulnerable against redundant low-quality images, *i.e.* when the number of low quality images is high, it could be easily distracted. The first 6 images in the IJB-S video are nearly identical images, but overall they have even a larger weight than the two higher quality faces. Similarly, in the PaSC case, the last 5 images almost contain the same information, but overall they gain 41% attention. In comparison, REAN is more robust against redundancy.

5.5. Quantitative Analysis on IJB-S

Table 2 reports identification results on IJB-S dataset. One of the comparisons is between the proposed method and previous work on IJB-S [11]. It is clear that the proposed approach achieves the best performance on most of the five protocols. In particular, the proposed approach outperforms C-FAN 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 REAN 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 may not be as accurate as surveillance frames. Therefore, for booking images, we use average pooling instead of quality aggregation. In comparison with the three baseline methods, REAN 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 gener-

alize to the booking images. Since the range of co-domain of *LSTM* is larger than that of REAN, the distribution of the output vector from *LSTM* can be very different from original feature vector, so the matching between surveillance videos and still images leads to poor performance. *Quality-Pool* achieves similar performance as C-FAN, since both of the approaches use component-wise attention scheme. Both *QualityPool* and the proposed REAN outperform average pooling of the base CNN features.

5.6. 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. REAN outperforms the three baseline methods in terms of average verification accuracy. The performance of REAN is slightly higher than previous approaches such as NAN [47] and C-FAN [11]. 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 have obvious advantages over other state-of-the-art approaches.

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 REAN 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 per-

Table 2: Performance comparisons on IJB-S dataset.

Test Name	Method	Closed-set (%)			Open-set (%)	
		Rank-1	Rank-5	Rank-10	1 % FPIR	10 % FPIR
Surveillance-to-still	C-FAN [11]	50.82	61.16	64.95	16.44	24.19
	AvgPool	50.80	59.60	64.86	11.60	20.84
	LSTM	2.90	17.46	38.10	0.12	2.80
	QualityPool	51.61	62.78	66.39	17.33	23.91
	REAN	**57.59**	**64.07**	**68.39**	**20.06**	**28.44**
Surveillance-to-booking	C-FAN [11]	53.04	62.67	66.35	27.40	29.70
	AvgPool	50.82	59.73	64.25	19.19	25.43
	LSTM	4.49	16.40	32.81	1.31	15.97
	QualityPool	52.66	62.95	65.86	25.60	31.83
	REAN	**58.52**	**65.21**	**68.62**	**31.25**	**33.10**
Multi-view	C-FAN [11]	96.04	99.50	99.50	70.79	85.15
	AvgPool	96.53	98.51	99.01	66.33	82.17
	LSTM	4.95	19.80	39.61	3.21	23.56
Surveillance-to-booking	*QualityPool*	97.03	99.50	99.50	75.62	86.21
	REAN	**98.02**	**99.50**	**99.50**	**76.24**	**93.07**
Surveillance-to-Surveillance	C-FAN [11]	10.05	17.55	21.06	0.11	0.68
	AvgPool	7.71	14.34	18.95	0.08	0.57
	LSTM	11.13	21.07	23.52	0.08	0.20
	QualityPool	5.69	9.43	13.00	0.09	0.34
	REAN	**21.96**	**33.79**	**36.30**	**0.21**	**0.97**
UAV	C-FAN [11]	7.59	**12.66**	20.25	0.00	0.00
	AvgPool	2.53	6.33	7.59	0.00	0.00
	LSTM	1.30	2.66	6.58	0.00	0.00
Surveillance-to-booking	*QualityPool*	7.59	10.85	13.92	0.00	1.04
	REAN	**8.06**	11.39	**23.92**	**3.13**	**3.13**

Table 3: Verification Performance on YouTube Face dataset, compared with baseline methods and other state-of-the-art methods.

Method	Accuracy (%)	Method	Accuracy (%)
EigenPEP [25]	84.8 ± 1.4	DeepFace [41]	91.4 ± 1.1
DeepID2+ [40]	93.2 ± 0.2	C-FAN [11]	96.50 ± 0.90
FaceNet [37]	95.52 ± 0.06	DAN [34]	94.28 ± 0.69
NAN [47]	95.72 ± 0.64	QAN [28]	96.17 ± 0.09
AvgPool	96.24 ± 0.96	*LSTM*	60.00 ± 2.81
QualityPool	96.38 ± 0.95	REAN	**96.60 ± 1.00**

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

Method	Control	Handheld
DeepO2P [30]	68.76	60.14
SPDNet [19]	80.12	72.83
GrNet [21]	80.52	72.76
Rao *et al.* [35]	95.67	93.78
TBE-CNN [8]	95.83	94.80
TBE-CNN + BN [8]	97.80	96.12
AvgPool	91.27	74.30
LSTM	3.07	1.28
QualityPool	96.48	92.39
REAN	96.52	94.63

formance compared to other state-of-the-art methods.

6. Conclusions

This paper addressed video face recognition for unconstrained surveillance systems. We propose a Recurrent Embedding Aggregation Network (REAN) that adaptively predicts context-aware quality vectors for each deep feature vector extracted by CNN face model. And each face in a video can perform efficient recognition algorithm by using a compact deep feature vector aggregated from REAN under the quality attention scheme. Experiments have been conducted on three video face datasets, i.e., IJB-S, YTF, and PaSC. Based on the recognition results, we empirically

demonstrate that the attention values provided by the proposed REAN is able to maintain the discriminative features while discard the noisy features by leveraging the context information in the video learned by LSTM. Our method shows clear advantages on video face benchmarks, including surveillance videos.

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, 2

[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. 5, 6

[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. 2

[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, 5

[5] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, 2010. 1, 2

[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. 2

[7] D. P. Dee. Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 2005. 3

[8] C. Ding and D. Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Trans. on PAMI*, 2018. 2, 8

[9] 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

[10] 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. ACM, 2016. 2

[11] S. Gong, Y. Shi, and A. K. Jain. Video face recognition: Component-wise feature aggregation network (c-fan). 2019. 2, 3, 6, 7, 8

[12] 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, 3

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

[14] 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, 6

[15] 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, 2

[16] 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

[17] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017. 5

[18] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, 2011. 2

[19] Z. Huang and L. Van Gool. A riemannian network for spd matrix learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 8

[20] 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. 2

[21] Z. Huang, J. Wu, and L. Van Gool. Building deep networks on grassmann manifolds. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 8

[22] 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. 2, 4, 5

[23] 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. 2, 4

[24] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, 2003. 2

[25] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-pep for video face recognition. In *ACCV*, 2014. 8

[26] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 1, 4, 6

[27] 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. 1, 2

[28] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017. 1, 2, 3, 6, 8

[29] 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. 2

[30] E. Mohagheghian. *An application of evolutionary algorithms for WAG optimisation in the Norne Field*. PhD thesis, Memorial University of Newfoundland, 2016. 8

[31] 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. 2

[32] 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

[33] 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. 2

[34] Y. Rao, J. Lin, J. Lu, and J. Zhou. Learning discriminative aggregation network for video-based face recognition. In *ICCV*, 2017. 1, 8

[35] Y. Rao, J. Lu, and J. Zhou. Attention-aware deep reinforcement learning for video face recognition. In *ICCV*, 2017. 1, 2, 8

[36] Y. Ren, K. Anderson, K. Iftekharuddin, P. Kim, and E. White. Pose invariant face recognition using cellular si-

multaneous recurrent networks. In *International Joint Conference on Neural Networks*, 2009. 2

[37] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1, 8

[38] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *ECCV*, 2002. 2

[39] 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. 1

[40] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, 2015. 1, 8

[41] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1, 8

[42] 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, 2

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

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

[45] W. Xie, L. Shen, and A. Zisserman. Comparator networks. In *ECCV*, 2018. 1

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

[47] 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. 1, 2, 3, 4, 6, 7, 8

[48] M. Yang, P. Zhu, L. Van Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. In *IEEE FG*, 2013. 2

[49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016. 6

[50] 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

[51] 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

[52] Y. Zhong, R. Arandjelović, and A. Zisserman. Ghostvlad for set-based face recognition. *arXiv:1810.09951*, 2018. 1