Geometry Guided Feature Aggregation in Video Face Recognition

Baoyun Peng[1], Xiao Jin[2], Yichao Wu[2], Dongsheng Li[1]*,
[1] Natioan University of Defense Technology, [2] Sensetime Group Limited,
{pengbaoyun13, dsli}@nudt.edu.com jinxiaocuhk@gmail.com wuyichao@sensetime.com

Abstract

Video-based face recognition has attracted a significant amount of research interest in both academia and industry due to its wide applications such as surveillance and security. Different from image-based face recognition, abundant information, extracted from a series of frames in a video, would contribute a lot to successful recognition. In other words, the key to improving video face recognition capability is aggregating and integrating profuse information within a video. Existing methods of feature aggregation across frames narrowly focus on the importance of a single frame, while ignoring the geometric relationship among frames in feature space. In this work, we present a geometry-based feature aggregation method rather than a better recognition model. It considers not only the importance of each frame but also the geometric relationship among frames in feature space, which yields more distinguishing video-level representation. Extensive evaluations on IJB-A and YTF datasets indicate that the proposed aggregation method considerably outperforms other feature aggregation methods.

1. Introduction

The performance of automatic face recognition (FR) has been considerably improved in recent years, mainly owing to the combination of algorithms based on deep neural networks and large scale labeled face data. Remarkably, on the representative academic benchmarks LFW [16], several FR methods [30, 25, 28] have even surpassed human for face verification. Different from image-based face recognition, in video-based FR, much more information about identities can be extracted from image sequences. However, video faces usually suffer from low quality due to unconstrained variations of poses, illuminations, blurriness and etc. These factors result in larger intra-class variance and sharply degrade the performance of face recognition. Hence, the key point to improve video face recognition performance is aggregating abundant information across frames, meanwhile discarding noise as well as maintain essential information.

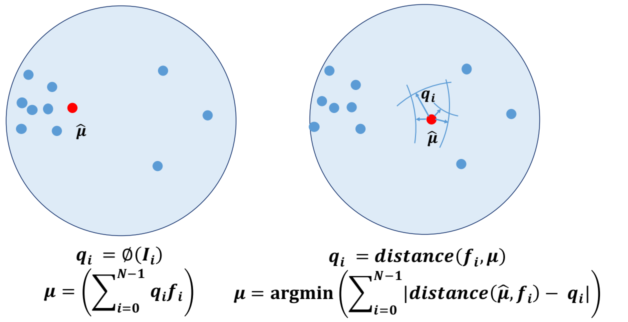
q_i
$\hat{\mu}$
$\hat{\mu}$
$q_i = \emptyset(I_i)$
$\mu = \left(\sum_{i=0}^{N-1} q_i f_i\right)$
$q_i = distance(f_i, \mu)$
$\mu = \mathbf{argmin}\left(\sum_{i=0}^{N-1} |distance(\hat{\mu}, f_i) - q_i|\right)$

Figure 1. The difference between quality based and geometry based feature aggregation strategy. Quality based method only considers the importance of each single frame, while GFA considers both the quality of image and its feature's geometric location. The equations under the figures indicate the way to figure out the reference feature from a group of features.

Many efforts have been proposed to effectively aggregate information to form a distinguishing and discriminative video face representation [3, 20, 17, 6, 8]. Feature aggregation is an effective method to integrate multiple features into a compact discriminative representation. For instance, Max Pooling and Average Pooling strategies are two naive aggregation strategies, which are commonly adopted in [29, 25]. A smarter strategy is to weigh and combine the image-level features, shown in [35, 6, 27, 21]. Usually, these methods aim at learning the weights of features from images or features themselves according to face image quality. Since image quality is a criterion represents the importance of each single frame in these methods, it is appropriate to call them quality based method. Here, we should point out that these quality-based methods fail to consider the relation among frames in feature space, and such neglect would lead to an incompetent aggregated feature.

We argue that the goal of feature aggregation is to make the aggregated feature of a subject closer to the center of its distribution in feature space. The reason is that this kind of aggregation or representation is more discriminative. How-

*Corresponding author.

ever, making the aggregated feature closer to the center is difficult since the distribution is unknown. To solve this problem, we proposed a geometry-based feature aggregation (GFA) method to generate a more discriminative aggregated feature. The details of the proposed method is that: first, we define the feature of a reference face image (under natural pose and illumination, no distortion) as the center in feature space since these images are easier to recognition, and utilize the geometric distances between frames' features and this center as the quality label to train a quality assessment(QA) model; second, such a QA model is adopted to predict the geometric distance between the image feature and its distribution center; finally, the center, calculated according to predicted geometric distances, is regarded as the final aggregated feature for that subject. Figure 1 shows the flowchart of the proposed method. Experimental results on two representative video face datasets show the effectiveness of the proposed GFA. The contributions of this paper are summarized as follows:

- we propose a novel feature aggregation method based on geometry relation in feature space;
- we design an image quality metric based on geometric distance from the center in feature space;
- we demonstrate the effectiveness of the proposed method in seeking a more discriminative video representation.

The remaining part is organized as follows. We first review related works on video-based face recognition and feature aggregation. Next, a novel geometry-aware quality assessment network is presented in details. Then, we describe the GFA in detail. Furthermore, comprehensive evaluation for feature aggregation on IJB-A and YTF datasets is illustrated in the experimental section. Finally, conclusion, remarks, and discussions end up the whole paper.

2. Related Work

Different from face recognition based on a single image, much more information about identity can be extracted from different images that are captured under multiple views. However, face images in videos usually suffer from low-quality. Many factors jointly cause this low-quality: blur, low-resolution, large-variation illumination, pose, etc. These factors result in larger intra-class variance that causes sharp degradation on the performance of face recognition. Consequently, the key to increasing the accuracy of video face recognition is to aggregate useful identity information and discarding noises at the same time.

Video-level representation with metric learning Many works on video or image set based face recognition attempt to model images' set as convex hull [3, 15] or subspace [36, 17]. So these authors assume that a Grassmann manifold or Hilbert space can successfully model samples distribution. They use manifold similarity metric to gain video-level representation [36, 22, 17]. In addition, several works attempted to fuse local features to build video feature representation [19, 20, 34]. Classical works include VLAD [18], PEP-Eigen [20] and Fisher Vector Faces [24]. Admittedly, these methods limit to work under constrained settings.

Feature-level aggregation Average Pooling or Max Pooling has been wildly applied in most popular recognition methods [25, 20, 5]. However, treating the difference between frames with indifference, by using Average Pooling or Max Pooling, would have negative effects. Therefore, a more reasonable aggregating strategy would fuse features according to the importance or the quality of each feature [35, 21], which shows superiority to Average Pooling or Max Pooling strategies. Along with another axis, temporal dynamic information among frames in a video is considered to guide aggregation [9, 27, 14]. For example, Rao et al. proposed an attention-aware method to explore the temporal information and view the attention problem as a Markov decision process. To train such an attention model, a deep reinforcement learning framework was proposed in their work.

Image-level aggregation Another strategy to aggregate information across frames is performing aggregation on image-level to synthesize one or few discriminative frontal face images [26, 23, 31, 38, 39] for face recognition. Usually, these methods require a complex model, mostly generative adversarial network (GAN) [10], to accomplish the synthesizing procedure, which largely limits its application scope. Besides, aggregation on matching results are not desirable, especially for large-scale recognition due to its high space complexity. Compared with two methods mentioned above, feature aggregation, which fuses multiple features into a compact discriminative video-level representation, is a more advisable approach that can be easily incorporated into currently existing face recognition systems.

Note that, in this paper, we mainly focus on the dataset containing order-less images, including both video based and template based datasets. Different from quality-based feature aggregation which tries to learn face image quality from images or features and designates quality as the weight of feature, our geometry based aggregation can directly estimate the center of the distribution, and adopt this center as the final aggregated feature.

3. Proposed method

Figure 2 illustrates the overall pipeline of the proposed method. The overall framework consists of three modules: quality assessment (QA), feature extraction, and feature aggregation. The QA module takes the video frames as input and predicts the quality score, which indicates the geomet-

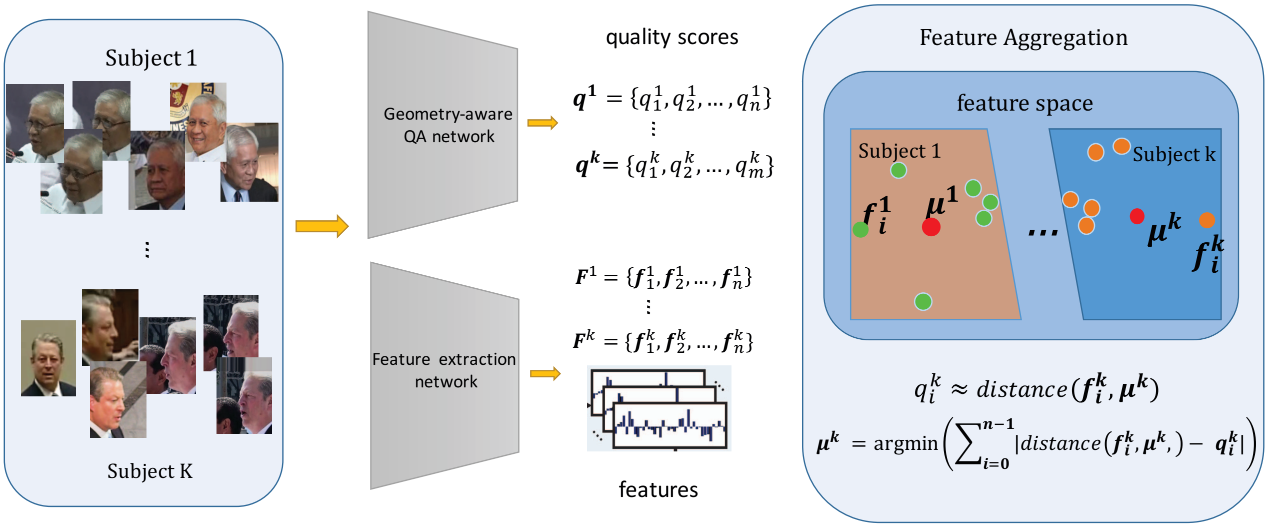
Subject 1
⋮
Subject K
Geometry-aware QA network
Feature extraction network
quality scores
$q^1 = \{q_1^1, q_2^1, \ldots, q_n^1\}$
⋮
$q^k = \{q_1^k, q_2^k, \ldots, q_m^k\}$
$F^1 = \{f_1^1, f_2^1, \ldots, f_n^1\}$
⋮
$F^k = \{f_1^k, f_2^k, \ldots, f_n^k\}$
features
Feature Aggregation
feature space
Subject 1
Subject k
f_i^1
μ^1
…
μ^k
f_i^k
$q_i^k \approx distance(f_i^k, \mu^k)$
$\mu^k = \text{argmin}\left(\sum_{i=0}^{n-1} |distance(f_i^k, \mu^k,) - q_i^k|\right)$

Figure 2. The overall framework of quality and geometry guided feature aggregation. A single image is the input of both geometry-aware QA network and feature extraction network. The results of the two networks are the geometry-based quality and feature representation of this image. The final feature aggregation process is aimed at finding out the reference feature according to the equation above.

ric distance between the image feature and the distribution center. Besides, the QA module also plays a role in frame selection. More specifically, the QA module will filter out those frames with low quality. Feature extraction module extracts feature from remaining frames. Since we assume that the quality score of each frame indicates the distance between current frame feature and the distribution center in feature space, it is easy to calculate the center according to the geometry relationship between features and the center when sufficient frames are provided. Therefore, feature aggregation module is designed to estimate the final feature by integrating the generated features according to the output of the QA module.

3.1. Geometry-aware QA Network

In video-based FR, face images are captured under a circumstance combined with large variation. Such circumstance results in much dross frames. Thus, discarding those low-quality samples and selecting the essential information can enhance the robustness of recognition result. [21, 1].

In this paper, we present a geometry-aware quality metric to assess face images' quality. In this work, matching score, to be more specifically cosine similarity, between current feature and reference feature is considered to be the quality of each frame. It should be pointed out that an approximation of the center point of a person's features in feature space is adopted to be the reference feature mentioned above. In addition, to get the reference feature, we suppose that a face image, which is under frontal pose, uniform illumination circumstance, and freed from disturbance, would possess a feature locating at the vicinity of the center point of that person's features, and we can integrate several features extracted from such kind of face images to generate the reference feature.

There are two reasons that lend credible support for our assumption: On the one hand, this quality metric gives out the matching results directly, and we can easily select those frames possessing a good matching result; On the other hand, this quality metric represents the geometric distance between current feature and reference feature in feature space, and a valid feature aggregation can be successfully guided by this geometric distance.

Consider a frame set $\{X_i | i = 1, 2, ..., n\}$ in a video sequence, and the embedding features $\mathbf{f_i} \in R^d$ are extracted from image X_i. Ordinarily, each feature f_i can be regarded as a point in feature space. Additionally, in this paper, we denote reference feature as μ.

Thus, the quality score of X_i can be computed as follows:

$$q_i = d(\mathbf{f}_i, \mu) \tag{1}$$

where $d(\mathbf{f}_i, \mu)$ represents the distance between $\mathbf{f}_i$ and μ. To keep correspondence with the recognition model, cosine similarity is naturally adopted as the distance metric in this work. Inspired by the success of deep CNN in various vision tasks, we use DCNN architecture to predict the quality corresponding to raw face image automatically.

3.2. Geometry Guided Feature Aggregation

Different from image-based face recognition, video-based face recognition conduct identification and verification on videos, which contain multiple frames under various

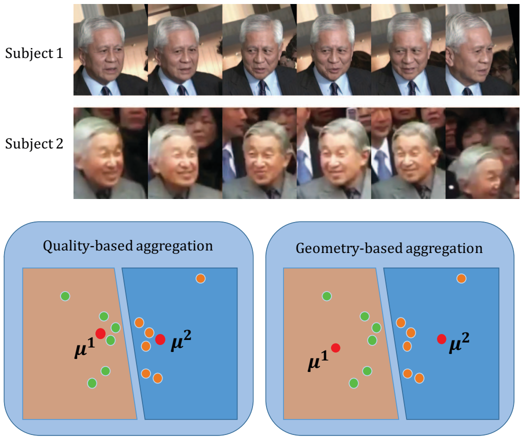

Figure 3. Geometry-based aggregation outstrips quality-based aggregation when the distribution is uneven because the geometry-based metric allows it to overcome the unbalance distribution. Reference features given by quality-based aggregation is much more closer than their counterparts given by geometry-based aggregation, which means a inferior representation.

conditions(e.g., pose, illumination, and blur). Compared with a single image, a video provides us with abundant information across frames. Therefore, how to aggregate information across frames to alleviate the problem caused by noise within a video and how to get more valuable and more effective representation for robust recognition against variations are a crucial issue for promoting FR performance. Feature aggregation is one of the effective ways of integrating information across multiple frames to form a video-level representation. Obtaining a valuable group of weight to weigh the corresponding features should be the core of the feature aggregation algorithm. In this section, we show that the feature aggregation problem can be formularized as a minimum squared error problem.

Let's say that $\chi = \{X_1, X_2, ..., X_n\}$ stands for a sequence of frames in a video, and $\mathbf{f}_i$ represents the corresponding normalized feature vector ($\|f_i\|_2 = 1$) for image X_i. The image set in a template or video is viewed as an approximation to a convex hull, just like what [3, 21, 35] assume. Then, the aggregated feature μ can be computed from features $\{f_i | i = 1, 2, n\}$ in a image set as the following equation illustrates:

$$\mu = \sum_{i=1}^{n} \alpha_i \mathbf{f}_i \tag{2}$$

Here, α_i denotes the weight of i_{th} feature of a subject, and satisfies the restriction: $\sum_{i=1}^{n} \alpha_i = 1$. The aggregated feature μ is in the affine hull spanned by features samples $\{f_1, f_2, ..., f_n\}$.

A widely adopted method to obtain the weight α_i is averaging the aggregation by setting $\alpha_i \equiv \frac{1}{n}$ [25, 4]. Beyond that, in [35], a two-cascaded attention blocks is enrolled to learn the importance $\mathbf{e_i}$ of each feature, and the weight is computed as $\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)}$. Furthermore, in [21], authors proposed a quality-aware feature aggregation method, which directly use the quality score as the weight of each feature, and the aggregated feature can be computed as follows: $\alpha_i = \frac{q_i}{\sum_{j=1}^{n} q_j}$. All the aforementioned method is based on the quality of a single image, no matter how the quality score is gotten. But in our geometry-guided method represents the distance between the current feature and its ideal center. Figure 3 shows the difference between aggregation strategies based on quality score and geometry relationship.

While quality-based aggregation strategies fail to take the distribution among frames into account, so they are more susceptible to unbalanced distribution. Consequently, quality-based aggregation would yield a reference feature vector that is close to the cluster of features. On the contrary, geometry-based aggregation method can avoid this phenomenon. As Figure 3 shows, features of one subject will lay together in the feature space when the frames in a video are lack of discrimination. Under such circumstance, geometry-based method rather than quality-based method can still give out a valuable reference feature: two subjects' reference features given by quality-based aggregation are much closer than their counterparts provided by our geometry-based method. So geometry-based aggregation would generate reference features that are more distinguishing.

Since our quality assessment module adopts a geometry-aware quality metric, we can get the following formulation:

$$q_i^{'} = \mathbf{f}_i^T \mu \approx \mathbf{f}_i^T \sum_{i=1}^{N} \alpha_i \mathbf{f}_i \tag{3}$$

where all the features $\mathbf{f}_i$ and reference feature μ are normalized to unit length.

Given a series of images in a video or image set, the quality of each image indicates the geometric distance between image feature and reference feature in a convex hull. The goal of feature aggregation is to fuse multiple features to a single feature. Based on Formula 3, optimization goal can be formularized as follows:

$$E(\alpha) = \frac{1}{2} \sum_{i=1}^{n} (\mathbf{f}_i^T \cdot \sum_{i}^{n} \alpha_i \mathbf{f}_i - q_i)^2 \tag{4}$$

$E(\alpha)$ represents the squared error between predicted quality score and ground truth. How to acquire $minE(\alpha)$ is a standard minimum squared error problem under the constraint ($\sum_{i=1}^{n} \alpha_i = 1$). To make the reference feature μ be more robust to outliers, we add L2 normalization $\alpha^T \alpha$ as regularization item and rewrite formula 4. To simplify the form, let

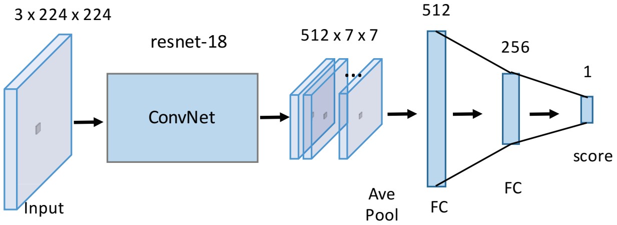

Figure 4. The details of the quality network. The quality assessment (QA) network is designed for estimating the geometric distance between the input image's feature and the ideal center point in the feature space of the same subject.

$F = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n] \in R^{d \times n}$, $\mathbf{q} = [q_1, q_2, ..., q_n]^T \in R^{n \times 1}$. Equation 4 can be rewritten as follows:

$$
\begin{aligned}
E(\alpha) &= \frac{1}{2}\sum_{i=1}^{n}(\mathbf{f}_i^T F\alpha - q_i)^2 + \frac{\lambda}{2}\alpha^T\alpha \\
&= \frac{1}{2}\alpha^T(F^T(\sum_{i=1}^{n}\mathbf{f}_i\mathbf{f}_i^T)F + \lambda I)\alpha - \mathbf{q}^T F^T F\alpha + \frac{1}{2}\mathbf{q}^T\mathbf{q}
\end{aligned}
\tag{5}
$$

Similar to [3], we introduce L and U coefficients to reduce affine hulls for controlling the looseness of the convex approximation. Finally, we can get the α by solving the following quadratically constrained quadratic program (QC) problem:

$$
\begin{aligned}
& E(\alpha) = \frac{1}{2}\alpha^T G\alpha + c^T\alpha + Constant \\
s.t. \quad & \sum_{i=1}^{n}\alpha_i = 1 \\
& \forall i \in 1, 2, ..., n, L \le \alpha_i \le U \\
opt. \quad & \arg\min_{\alpha} E(\alpha)
\end{aligned}
\tag{6}
$$

where $G = (F^T(\sum_{i=1}^{n}\mathbf{f_i}\mathbf{f_i^T})F + \lambda I)$ is a positive definite matrix, $c = F^T F\mathbf{q}$, λ represents the coefficient of L2 normalization, and $Constant = \frac{1}{2}\mathbf{q}^T\mathbf{q}$ is invariant to α. In this paper, we set the $U = 0.7$ and $L = 0$, $\lambda = 0.3$. There are many mature algorithms and tools to solve above QC problem. In this work, we choose the *quadprog* in matlab.

Not only does GFA focus on the importance of each frame, but also considers the geometric relation. To some extent, GFA can also be regarded as a effective method which can estimate the center of a set in approximated convex hull.

3.3. Implementation details

The details of quality network are shown in Figure 4.

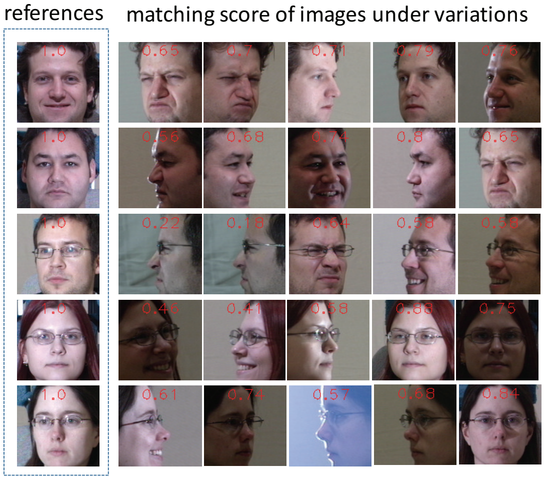

Figure 5. Part of images samples for training quality assessment network. The quality score is drawn on the image.

Multi-PIE [11] is used to train the QA network. Figure 5 shows the samples in Multi-PIE. We take the face with frontal pose, normal illumination as the reference, and employ cosine similarity between image and corresponding reference in feature as the quality score.

ResNet-18 [13] is adopted to be the base model of QA network,and the initial learning rate is 0.001, batch size is set to 256. All of the input images are aligned and resize to 96 × 96. L1 Loss is used to train such a regression network due to its statistical property is superior to L2 Loss in literature. The loss function is shown as following:

$$
L = \frac{1}{N}\sum_{i=1}^{N}\|\phi(X_i, \mathbf{f}_i; \theta) - q_i\|_1 \tag{7}
$$

where N denotes the number of face images in training dataset and $\phi(X, f; \theta)$ denotes the quality assessment model θ.

4. Experiments

We evaluate the proposed method on two representative videos (and template) face datasets: IJB-A and YTF, and compare it with other state-of-the-art video face recognition methods, especially with other feature aggregation methods. What we should point out is that different from other aggregation methods, none of the data in IJBA or YTF are used as the training dataset.

4.1. Experiment settings

IJB-A dataset consists of 5,712 images and 2,085 videos, which are captured from 500 subjects under unconstrained conditions. The IJB-A challenge aims at evaluating FR and verification on templates containing order-less images set or videos. We follow the IJB-A protocol: 1:1 verification

Table 1. Performance of GFA, average, quality pooling under different number of frames.

method	IJBA 1:1 verification				IJBA 1:N Identification			
	FAR=0.0001	FAR=0.001	FAR=0.01	FAR=0.1	FAR=0.01	FAR=0.1	Rank-1	Rank-10
average / 5frs	0.629 ± 0.028	0.904 ± 0.029	0.948 ± 0.017	0.968 ± 0.007	0.534 ± 0.338	0.929 ± 0.013	0.949 ± 0.015	0.975 ± 0.011
average / 5frs	0.629 ± 0.028	0.904 ± 0.029	0.948 ± 0.017	0.968 ± 0.007	0.534 ± 0.338	0.929 ± 0.013	0.949 ± 0.015	0.975 ± 0.011
quality / 5frs	0.658 ± 0.025	0.905 ± 0.029	0.948 ± 0.016	0.968 ± 0.007	0.545 ± 0.344	0.931 ± 0.015	0.950 ± 0.014	0.975 ± 0.011
GFA / 5frs	**0.748** ± **0.018**	**0.913** ± **0.034**	**0.952** ± **0.018**	**0.968** ± **0.009**	**0.713** ± **0.222**	**0.946** ± **0.015**	**0.956** ± **0.021**	**0.976** ± **0.008**
average / 10frs	0.629 ± 0.029	0.92 ± 0.021	0.956 ± 0.012	0.971 ± 0.005	0.618 ± 0.3-6	0.938 ± 0.014	0.954 ± 0.972	0.975 ± 0.008
quality / 10frs	0.654 ± 0.027	0.921 ± 0.020	0.957 ± 0.012	0.971 ± 0.005	0.644 ± 0.280	0.939 ± 0.015	0.955 ± 0.016	0.972=0.008
GFA / 10frs	**0.833** ± **0.012**	**0.948** ± **0.016**	**0.973** ± **0.005**	**0.982** ± **0.007**	**0.917** ± **0.037**	**0.964** ± **0.007**	**0.971** ± **0.008**	**0.983** ± **0.005**
average / 30frs	0.837 ± 0.012	0.95 ± 0.013	0.973 ± 0.007	0.983 ± 0.006	0.912 ± 0.041	0.964 ± 0.009	0.972 ± 0.007	0.984 ± 0.006
quality / 30frs	0.852 ± 0.010	0.951 ± 0.006	0.973 ± 0.006	0.983 ± 0.006	0.919 ± 0.035	0.965 ± 0.009	0.972 ± 0.007	0.984 ± 0.007
GFA / 30frs	**0.867** ± **0.086**	**0.951** ± **0.006**	**0.973** ± **0.006**	**0.984** ± **0.006**	**0.929** ± **0.030**	**0.966** ± **0.009**	**0.972** ± **0.008**	**0.985** ± **0.004**
max	0.555 ± 0.034	0.889 ± 0.027	0.933 ± 0.014	0.957 ± 0.010	0.477 ± 0.376	0.910 ± 0.013	0.939 ± 0.013	0.967 ± 0.013
average / all	0.840 ± 0.08	**958** ± **0.004**	0.970 ± 0.008	**0.986** ± **0.004**	0.883 ± 0.07	0.960 ± 0.013	0.971 ± 0.012	0.985 ± 0.010
quality / all	**0.8722** ± **0.07**	0.956 ± 0.004	**0.972** ± **0.008**	0.983 ± 0.004	**0.964** ± **0.039**	**0.974** ± **0.008**	**0.971** ± **0.013**	**0.985** ± **0.004**

and 1:N identity, two tasks. YTF dataset is a video based dataset with 3,425 videos belonging to 1,595 different subjects. There is only 1:1 verification task in YTF. We adopt the PolyNet [37] as our recognition model, and trained it on IMDb-Face [32] and MS-Celeb-1M [12]. In our experiment, PolyNet [37] serves as the base model of the recognition network. All of the input images are aligned and resized to 224. We train the recognition network on IMDB dataset [32].

4.2. Evaluation for Feature Aggregation

We conduct extensive experiments about different aggregation strategies, including Max Pooling, Average Pooling, quality-based, and GFA. All the methods are evaluated on both IJB-A and YTF dataset. Table 1 shows the results of three different aggregation strategies under different number of images — 5, 10, 30. A conclusion that the proposed GFA method is much better than both quality-based and Average Pooling aggregation strategies would be evident. Another safe conclusion drawn from Table 1 is that using more images to fuse leads to higher performance for both quality pool and GFA. However, the performance of Average Pooling degrades when fusing all images in a template. A tenable explanation is that the dross (images that are too hard to be recognized) would have a negative influence on the aggregated feature, the representation of discriminative information. But such hurts of dross can be mitigated by removing images with low quality scores predicted by the QA network.

The results on YTF are shown in Table 2. GFA method outperforms ADRL by reducing 24% error ratio and incredibly achieving 97.2% accuracy, and it becomes the result with single model on YTF in publication, to our best knowledge. Similarly, as Table 3 shows, we evaluate different aggregation strategies with different number of images on YTF. From the result illustrated, we can find that the performance of quality-based method is almost the same as the one adopting Average Pooling and both of them are inferior to the result of GFA. GFA outperforms them by reducing the FRR from 0.064 to 0.053 at FAR=0.01, and 0.037 to 0.031 at FAR=0.1.

Table 2. Evaluation on YTF dataset. The comparing methods are listing as following: Centerloss [33], FaceNet [28], NAN [35], ADRL [27], QAN [21].

method	acc (%)	auc (%)
centerloss	94.9	
FaceNet	95.12+0.39	
NAN	95.72+0.64	98.8
ADRL	96.52+0.54	
QAN	96.17+0.09	99.14
GFA	**97.2**	

Table 3. Evaluation for different aggregation strategies with different number (5, 10, 30) of frames to fuse on YTF dataset.

method	FRR@FAR=0.01	FRR@FAR=0.1	ACC
max	0.072 ± 0.032	0.044 ± 0.024	0.96
average / 5frs	0.072 ± 0.032	0.044 ± 0.024	0.960
quality / 5frs	0.072 ± 0.032	0.044 ± 0.024	0.960
GFA / 5frs	**0.063** ± **0.035**	**0.040** ± **0.024**	**0.963**
average / 10frs	0.064 ± 0.036	0.037 ± 0.029	0.965
quality / 10frs	0.064 ± 0.036	0.037 ± 0.029	0.965
GFA / 10frs	**0.053** ± **0.030**	**0.031** ± **0.023**	**0.969**
average / 30frs	0.052 ± 0.020	0.030 ± 0.018	0.97
quality / 30frs	0.052 ± 0.020	0.030 ± 0.018	0.969
GFA / 30frs	**0.047** ± **0.019**	**0.030** ± **0.018**	**0.971**

4.3. Qualitative and Quantitative Evaluation

Plainly, the quality score is regarded as the importance in quality based method: higher quality score means higher weight in feature aggregation. While in GFA, weights are modeled as the solution of a constrained MSE problem to ensure that the aggregated feature is close to reference point enough in feature space. Sequentially, GFA generates a more discriminative feature. We visualize the images and corresponding scores predicted by QA network in Figure 6 to explain the function of QA network. Figure 6 shows different instances in 4 templates. All images in a template are sorted according to its quality score from high to low. Scrutiny the images shown in Figure 6 and we can find the difference of weight assignment between single image quality based method and geometry guided method: in quality based method, images that possess a good quality would be assigned a high weight and weigh more in generating the reference feature, while in geometry guided method, a group of images that loo Similar would be assigned a relatively low weight. It is evident that features extracted from similar images would locate close in feature space. In this way, the group of images assigned low weight serves as a single image with high weight. Therefore, the geometry guided method can successfully avoid generating a reference feature laying in the vicinity of a cluster of features. In other words, GFA can ensure the aggregated feature shall be freed from the influence of biased dense regions in feature space.

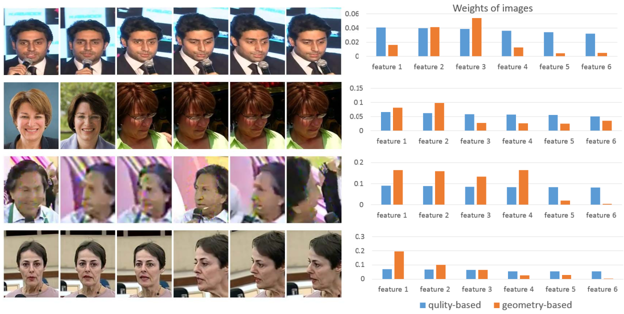

Figure 6. Instances of different templates in IJB-A. Each row shows six images sorted according to its quality scores from high to low. The histograms on the right represent the distribution of weights of the quality based method (blue) and GFA (orange). Images looked similar would be assigned a relatively lower weight in GFA.

In order to evaluate the effectiveness of QA network quantitatively, we test it on IJBA 1:1 and 1:N task. The results are shown in Table 4. Conclusion can be drawn from Table 4 that QA network does considerably promote the performance on recognition.

Table 4. Performance on the Arcface [7] model and PolyNet [37] model with/without QA. Arcface model is trained on VGGface2 dataset [2], and the PolyNet model is trained on IMDB dataset [32].

method	IJBA 1:1		IJBA 1:N
	FAR=0.01	FAR=0.001	Rank-1
Arcface w/o QA	0.598 ± 0.058	0.477 ± 0.084	0.789
Arcface w/ QA	**0.836 ± 0.019**	**0.779 ± 0.022**	**0.920**
PolyNet w/o QA	0.668 ± 0.038	0.587 ± 0.074	0.824
PolyNet w/ QA	**0.933 ± 0.014**	**0.889 ± 0.027**	**0.939**

5. Conclusions and Discussions

In this paper, we present a geometry guided feature aggregation method for video-based face recognition. Image quality and geometric relations between frames in feature space are fully considered in this work. By defining a geometry-aware quality metric, which regards the distance between current image feature and reference feature as quality score, we find that the feature aggregation problem can be formularized as a constrained minimum squared error (MSE) problem within an approximated convex hull. To solve such a constrained MSE problem, we convert it to a quadratically constrained quadratic program. In experiments, on two representative datasets, GFA method outperforms all of other state-of-the-art feature aggregation methods. Besides, another impressing advantage of GFA is that it does not require training recognition network again. In other word, GFA possesses high adaptability, which allows us to applied GFA to any general recognition models.

References

[1] Lacey Best-Rowden and Anil K Jain. Automatic face image quality prediction. *arXiv preprint arXiv:1706.09887*, 2017.

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE FG*, pages 67–74, 2018.

[3] H Cevikalp and B Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, 2011.

[4] Juncheng Chen, Rajeev Ranjan, Amit Kumar, Chinghui Chen, Vishal M Patel, and Rama Chellappa. An end-to-end system for unconstrained face verification with deep convolutional neural networks. *ICCV*, pages 360–368, 2015.

[5] Aruni Roy Chowdhury, Tsung Yu Lin, Subhransu Maji, and Erik Learned-Miller. One-to-many face recognition with bilinear cnns. In *WACV*, pages 1–9, 2016.

[6] Nate Crosswhite, Jeffrey Byrne, Chris Stauffer, Omkar Parkhi, Qiong Cao, and Andrew Zisserman. Template adaptation for face verification and identification. In *IEEE FG*, pages 1–8. IEEE, 2017.

[7] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv:1801.07698*, 2018.

[8] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, Song Shi, and Stefanos Zafeiriou. Lightweight face recognition challenge. In *ICCV*. IEEE, 2019.

[9] Basura Fernando, Efstratios Gavves, Oramas M Jos, Amir Ghodrati, and Tinne Tuytelaars. Rank pooling for action recognition. *TPAMI*, 39(4):773–787, 2017.

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *NIPS*, 3:2672–2680, 2014.

[11] R Gross, I Matthews, J Cohn, and T Kanade. Multi-pie. In *IEEE FG*, pages 1–8, 2010.

[12] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for

large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[14] Xiaowei Hu, Lei Zhu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Recurrently aggregating deep features for salient object detection. In *AAAI*, 2018.

[15] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, pages 121–128, 2011.

[16] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[17] Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Projection metric learning on grassmann manifold with application to video based face recognition. In *CVPR*, pages 140–149, 2015.

[18] Herve Jegou, Matthijs Douze, Cordelia Schmid, and Patrick Perez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311, 2010.

[19] Haoxiang Li, Gang Hua, Zhe Lin, and Jonathan Brandt. Probabilistic elastic matching for pose variant face verification. In *CVPR*, pages 3499–3506, 2013.

[20] Haoxiang Li, Gang Hua, Xiaohui Shen, Zhe Lin, and Jonathan Brandt. *Eigen-PEP for Video Face Recognition*. CVPR, 2013.

[21] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. pages 4694–4703, 2017.

[22] Jiwen Lu, Gang Wang, Weihong Deng, Pierre Moulin, and Jie Zhou. Multi-manifold deep metric learning for image set classification. In *CVPR*, pages 1137–1145, 2015.

[23] Tran Luan, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, pages 1283–1292, 2017.

[24] Omkar M. Parkhi, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. A compact and discriminative face track descriptor. In *CVPR*, pages 1693–1700, 2014.

[25] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, pages 41.1–41.12, 2015.

[26] Yongming Rao, Ji Lin, Jiwen Lu, and Jie Zhou. Learning discriminative aggregation network for video-based face recognition. In *CVPR*, pages 3801–3810, 2017.

[27] Yongming Rao, Jiwen Lu, and Jie Zhou. Attention-aware deep reinforcement learning for video face recognition. In *ICCV*, pages 3951–3960, 2017.

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[29] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, pages 1891–1898, 2014.

[30] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014.

[31] Luan Tran, Xi Yin, and Xiaoming Liu. Representation learning by rotating your faces. 2017.

[32] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. *arXiv preprint arXiv:1807.11649*, 2018.

[33] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016.

[34] Jian Xu, Cunzhao Shi, Chengzuo Qi, Chunheng Wang, and Baihua Xiao. Unsupervised part-based weighting aggregation of deep convolutional features for image retrieval. 2017.

[35] Jiaolong Yang, Peiran Ren, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *CVPR*, pages 1–8. IEEE, 2017.

[36] Meng Yang, Pengfei Zhu, Luc Van Gool, and Lei Zhang. Face recognition based on regularized nearest points between image sets. In *IEEE FG*, pages 1–7, 2013.

[37] Xingcheng Zhang, Zhizhong Li, Change Loy Chen, and Dahua Lin. Polynet: A pursuit of structural diversity in very deep networks. In *CVPR*, pages 3900–3908, 2017.

[38] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, and Junliang Xing. Towards pose invariant face recognition in the wild. In *CVPR*, 2018.

[39] Jian Zhao, Lin Xiong, Yu Cheng, Jianshu Li, Li Zhou, Yan Xu, Yi Cheng, Karlekar Jayashree, Sugiri Pranata, and Shengmei Shen. 3d-aided deep pose-invariant face recognition. In *IJCAI*, 2018.