Multi-View 3D Face Reconstruction with Deep Recurrent Neural Networks

Pengfei Dou and Ioannis A. Kakadiaris
Computational Biomedicine Lab, University of Houston
4800 Calhoun Road, Houston, TX 77004
{pdou,ikakadia}@central.uh.edu

Abstract

Image-based 3D face reconstruction has great potential in different areas, such as facial recognition, facial analysis, and facial animation. Due to the variations in image quality, single-image-based 3D face reconstruction might not be sufficient to accurately reconstruct a 3D face. To overcome this limitation, multi-view 3D face reconstruction uses multiple images of the same subject and aggregates complementary information for better accuracy. Though theoretically appealing, there are multiple challenges in practice. Of these challenges, the major one is that it is difficult to establish coherent and accurate correspondence among a set of images, especially when these images are captured in different conditions. In this paper, we propose a method, Deep Recurrent 3D FAce Reconstruction (DRFAR), to solve the task of multi-view 3D face reconstruction using a subspace representation of the 3D facial shape and a deep recurrent neural network that consists of both a deep convolutional neural network (DCNN) and a recurrent neural network (RNN). The DCNN disentangles the facial identity and the facial expression components for each single image independently, while the RNN fuses identity-related features from the DCNN and aggregates the identity specific contextual information, or the identity signal, from the whole set of images to predict the facial identity parameter, which is robust to variations in image quality and is consistent over the whole set of images. Through extensive experiments, we evaluate our proposed method and demonstrate its superiority over existing methods.

1. Introduction

Three dimensional face reconstruction from multi-view 2D images has been an active research topic for its great potential in many areas, such as facial animation, face synthesis, and facial recognition. Compared with single-view 3D facial shape reconstruction, multiple 2D images provide more information. However, it is a challenging task for existing methods to efficiently and effectively fuse information from multiple images.

Traditionally, most of the existing methods employ uncalibrated photometric stereo for 3D face reconstruction from a collection of facial images [22, 23, 26, 11, 12, 29]. Given multiple facial images of the same subject, uncalibrated photometric stereo first registers these images by using either facial landmark detection or optical flow to achieve dense pixel-wise correspondence. Then, matrix factorization is performed on the stack of registered facial images to recover the surface normal vector and the surface albedo at each pixel location. These methods suffer from several limitations. First, due to the smoothness and uniformity of human facial texture, it is difficult to establish dense and accurate correspondence at pixel level. Moreover, 2D registration cannot handle the out-of-plane rotation of human face in non-frontal facial images. As a result, uncalibrated photometric stereo is mostly constrained in recovering the 3D facial shape from near-frontal facial images. Second, uncalibrated photometric stereo assumes the Lambertian reflectance model of object surface, which is incapable of approximating complex scene illumination. Third, uncalibrated photometric stereo depends on a large number of images, usually more than 20 [22, 23, 11, 12], to achieve satisfying performance. This greatly increases the computational cost and constrains its applicability. Some other approaches employ structure from motion (SfM) to estimate the 3D facial shape from a collection of facial images [15, 25, 24, 14, 2, 31, 8]. Similar to uncalibrated photometric stereo, SfM first performs dense feature point detection on each facial image. Then, inter-image correspondence of the extracted feature points and extrinsic camera parameters is estimated via bundle adjustment [28]. Finally, multi-view stereo is employed to reconstruct the 3D point cloud. Compared with uncalibrated photometric stereo, SfM is robust to illumination and pose variations. However, it is difficult for SfM to reconstruct a dense 3D facial shape because the number of reliable feature points detected on the facial image is very limited. Moreover, SfM cannot handle non-rigid transform of the object, such as variations in facial expression, which is an important component of 3D face model-

ing. In addition, SfM is also computationally expensive and relies on a large number of facial images.

Several recent works have proposed employing deep neural networks (DNN) to solve the problem of 3D face reconstruction from a single 2D image [33, 10, 20, 21, 4]. In these works, based on a subspace facial shape model, each 3D face is represented as a model parameter vector and deep neural networks are trained to estimate the optimal parameter values from the 2D image. Although they achieved significant improvement over previous work [23, 12, 5] in reconstruction accuracy, they still cannot handle facial pose variation effectively. As a result, their performance degrades in 3D face reconstruction from non-frontal facial images with large pose variation. Moreover, they are focused on single-view 3D face reconstruction and have not provided a feasible solution to the multi-view 3D face reconstruction problem.

Our objective is to improve the accuracy of 3D face reconstruction by using multiple 2D images and improve the robustness to adverse factors including both illumination and pose variation. To this end, inspired by Dou *et al*. [4], we propose using a subspace model of 3D facial shape and employing a deep recurrent neural network (DRNN) to estimate the optimal model parameters from a small set of multi-view 2D images. Our deep recurrent neural network uses the deep convolutional neural network (DCNN) proposed by [4], from which the last fully-connected layer for estimating the facial identity parameters is removed. On top of this we add two Long-Short Term Memory (LSTM) layers [3] for the purpose of feature fusion and contextual information aggregation. The motivation behind our DRNN is that, in multi-view 3D face reconstruction, the subject identity is contextual information that is consistent over the set of 2D facial images. Compared with convolutional and fully-connected layers, the recurrent connection and the gating mechanism of the LSTM is capable of extracting the contextual information that is repeatedly passing through the neural network while filtering out the noise caused by adverse factors. Compared with SfM and uncalibrated photometric stereo, our approach is efficient and more robust to illumination and pose variation. Moreover, our approach is flexible to the number of available facial images and works well on very small image sets with only three or six 2D images, as demonstrated in the experimental evaluation. Compared with existing methods, our approach achieves significant improvement in both the accuracy and consistency of reconstructed 3D faces.

The rest of the paper is organized as follows. Section 2 reviews related work in the literature. Section 3 describes the details of our proposed method. Section 4 describes the implementation details and extensive experimental evaluations. Finally, Section 5 concludes the paper with a brief summary.

2. Related Work

Uncalibrated photometric stereo: Kemelmacher-Shlizerman *et al*. [12, 11] propose to factorize the observation using singular value decomposition (SVD) and take the rank-4 approximation to get the first order SH function and lighting coefficients. To resolve the ambiguity, they use a template 3D model to explicitly estimate the ambiguity parameters. One major limitation of the proposed method is that it cannot handle specularity reflection, which will corrupt the recovered SH subspace. To solve this problem, Snape *et al*. [26] propose a robust formulation for recovering the SH subspace and also estimating a sparse gross error term corresponding to these image artifacts. One major limitation of these methods is that they require many facial images to achieve a plausible reconstruction of the 3D facial model and rely heavily on the image registration accuracy. Moreover, they can only produce a 2.5D face surface with scale ambiguity. To solve these problems, Roth *et al*. [22] proposed a template deformation approach to reconstruct a 3D facial model from image collections. Instead of estimating a normal vector for each image pixel, they proposed to register a template 3D face to each 2D image through landmark-driven 3D warping and estimate only the normal vector of each 3D face vertex to deform the template 3D face. In spite of its improvement over previous work, due to the limitation of facial landmark detection, it still cannot handle facial images with large facial pose or challenging illumination. Moreover, it is computationally expensive and requires a large set of facial images to achieve acceptable reconstruction accuracy.

Structure-from-motion: Lin *et al*. [15] proposed using SfM for 3D face reconstruction from a set of five multi-view images. The Speeded Up Robust Features (SURFs) are extracted and matched in a pairwise manner to first estimate the fundamental matrix for each pair of images. Due to the smoothness and uniformity of human facial texture, only a sparse set of feature points can be extracted and registered. As a result, only a very coarse 3D face with sparse vertices can be reconstructed. Lee *et al*. [14] proposed applying SfM to reconstruct a coarse 3D face using facial landmarks detected on a set of three images. Then a template 3D face is warped to fit the coarse 3D face through thin-plate spline fitting. Similarly, Yang *et al*. [31] and Brunton *et al*. [2] also used SfM to reconstruct a coarse 3D face. The major limitation of SfM-based approaches is that they cannot handle the non-rigid transform of face caused by facial expression, which is an important component in 3D face modeling. Moreover, it is also challenging to establish accurate feature point correspondences among facial images captured in-the-wild.

3. Method

In this section, we elaborate on our proposed framework for multi-view 3D face reconstruction. We first provide an overview of our method. Then, we introduce the LSTM cell that builds the basis of our deep recurrent neural network. The network architecture is illustrated later with an in-depth analysis of its advantages in improving the robustness and consistency of 3D face reconstruction from multi-view images. Finally, we discuss the major training procedures, including network pre-training using both real and synthetic data and network fine-tuning.

3.1. Overview

Our goal is to reconstruct the 3D face of a subject using a set of facial images captured in different conditions, including time, camera pose, illumination condition, and resolution. To achieve this goal, we follow several recent works to use a subspace model for parametric 3D face representation and deep neural networks to approximate the nonlinear mapping from the image space to the optimal model parameters in the 3D facial shape subspace. Similar to [4, 33, 10, 20], the 3D facial shape subspace model consists of the mean 3D face $\bar{\boldsymbol{S}}$ and a set of shape and blendshape basis, $\boldsymbol{U}_d$ and $\boldsymbol{U}_e$. The shape basis $\boldsymbol{U}_d$ models the variation in neutral 3D facial shape, and thus is identity-specific, while the blendshape basis $\boldsymbol{U}_e$ models the variation in facial expression, and thus is not correlated with human identity. With different model parameters, different 3D faces can be reconstructed by:

$$\boldsymbol{S} = \bar{\boldsymbol{S}} + \boldsymbol{U}_d \cdot \boldsymbol{\alpha}_d + \boldsymbol{U}_e \cdot \boldsymbol{\alpha}_e \ , \tag{1}$$

where $\boldsymbol{S}$ is the target 3D face, $\boldsymbol{\alpha}_d$ is the facial identity parameter vector, and $\boldsymbol{\alpha}_e$ is the facial expression parameter vector. Given multiple images $\boldsymbol{I}^i$, $i \in 1, 2, \cdots, N$, of the same subject, the facial expression parameters $\boldsymbol{\alpha}_e^i$ of each image might be different, while the facial identity parameters $\boldsymbol{\alpha}_d^i$ should, ideally, be the same. Although several factors may affect the 3D facial shape, such as aging and weight, we assume that these factors remain similar when the facial images are captured. As a result, in solving the problem of multi-view 3D face reconstruction, the context that these 3D faces are from the same subject should be properly exploited and the identity specific contextual information should be properly extracted and preserved. Otherwise, the reconstructed 3D face will be inconsistent in terms of the facial identity.

To exploit, extract, and preserve the identity specific contextual information beneath the input 2D facial images, we propose to integrate a DCNN and a recurrent neural network (RNN) into a unified framework. The architecture of the proposed method is depicted in Fig. 1. The DCNN predicts for each single image the facial expression parameter vector $\boldsymbol{\alpha}_e^i$ and an identity-related feature vector $\boldsymbol{g}_d^i$, independently. Due to adverse factors, such as facial pose and illumination variation, $\boldsymbol{g}_d^i$ is noisy and inconsistent over the whole set of images. To extract the consistent facial identity parameter for face reconstruction, we add an RNN on top of DCNN to fuse the noisy features and aggregate the identity specific contextual information.

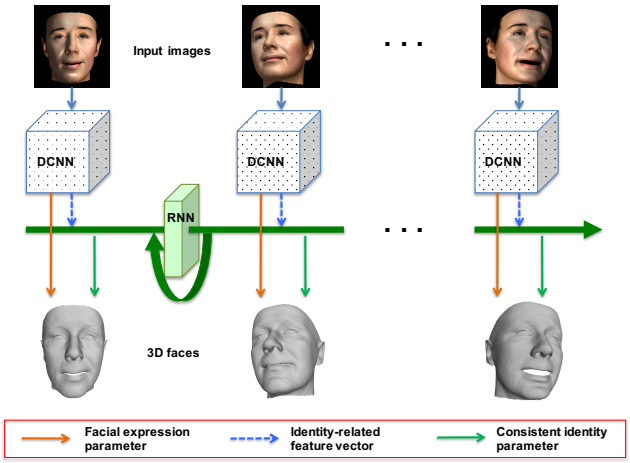


Figure 1: The depiction of the proposed deep recurrent neural networks. It consists of two components, a DCNN and an RNN. The DCNN processes each single image independently, while the RNN retains the contextual information from the whole image set.

The purpose of the DCNN is twofold. First, it approximates a nonlinear function that maps a single 2D image to deep features. Second, it disentangles the facial identity and the facial expression, allowing the consecutive RNN to exploit the identity specific contextual information. In our approach, we employ the DCNN framework proposed by Dou *et al*. [4], which consists of 13 convolutional layers for image feature extraction, a fusion-CNN for facial expression parameter estimation, and three fully-connected layers for facial identity parameter estimation. The last fully-connected layer is removed and replaced with the RNN that consists of two recurrent layers, each outputting a feature vector of 512 values, and a linear layer for estimating the facial identity parameters.

3.2. Recurrent Neural Networks with Long Short-Term Memory (LSTM) Cell

Recurrent neural networks refer to a kind of neural network with recurrent connections. Conventionally, RNN is proposed and designed to learn nonlinear mapping from a sequence of input $\boldsymbol{x}_t$ to a sequence of hidden states $\boldsymbol{h}_t$. The fundamental recurrent operation in RNN is achieved by the recurrent unit, such as the one depicted in Fig. 2. By recur-

sively feeding previous hidden state $\boldsymbol{h}_{t-1}$ into the repeating recurrent module, the recurrent unit is able to maintain a long-term memory of the contextual information. In application, such a recursive mechanism is demonstrated to be effective in extracting contextual information from the input sequence. However, the training of the vanilla recurrent unit is difficult due to the problem of exploding and vanishing gradients from long-term dependency tasks [17].

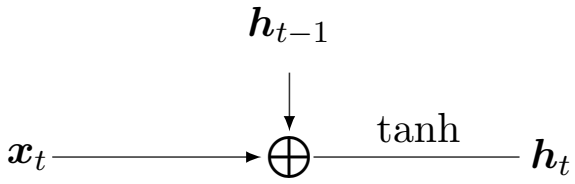


Figure 2: The RNN Conventional Graph (RCG) [6] depiction of a vanilla recurrent unit. In each time-step, previous hidden state is fed into the same unit through a recurrent connection. With the feedback loop, the recurrent unit is able to maintain a memory of the context.

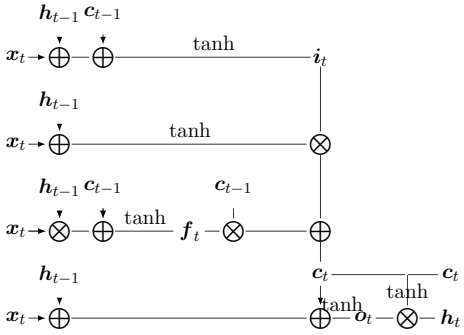


Figure 3: The depiction of the LSTM cell with a cell state and three modulating gates. The three modulating gates control how the cell state is updated and exposed.

To solve these problems, the LSTM unit, initially proposed by Hochreiter *et al*. [7] to explicitly tackle the long-term dependency problem, was revisited [3]. As illustrated in Fig. 3, the RCG of the LSTM cell, it introduces a new state, cell state $\boldsymbol{c}_t$, and three gates, namely an input gate i_t, an output gate o_t, and a forgot gate f_t, which control or modulate the information flow inside the LSTM cell. In principle, the input gate i_t controls the degree to which the new memory content is added to the cell state, the forgot gate f_t modulates the degree to which the existing memory is forgotten, and the output gate o_t controls the amount of memory content exposure. Compared with the recurrent unit, which overwrites its content at each time-step, the LSTM cell explicitly controls the behavior of memory updating via the introduced gates. As a result, the LSTM cell is able to decide whether to keep or forget the existing memory, being able to capture potential long-term dependencies. The modulating or gating mechanism introduced in the LSTM cell has been proven to be effective and achieved success in different areas, such as language modeling, video action recognition, and so on. It has also been observed that LSTM is more robust to noise in the data [19]. In this work, we use the LSTM cell to build our recurrent neural network, which is depicted in Fig. 4.

Our RNN consists of two recurrent layers, which are both LSTM cells, and a linear fully-connected layer. The input to the RNN is the identity-related feature $\boldsymbol{g}_d$ extracted by the DCNN. Given multiple images, the feature vector of each individual image $\boldsymbol{g}_d^i$ is fed into the RNN consecutively. The output of the RNN is a single facial identity parameter vector $\boldsymbol{\alpha}_d$, which is consistent over the whole set of images. The length of the feature vector $\boldsymbol{g}_d^i$ from the DCNN is 1,024. The outputs of the recurrent layers are feature vectors of 512 elements. The linear fully-connected layer on top of the LSTM cells outputs the facial identity parameter vector.

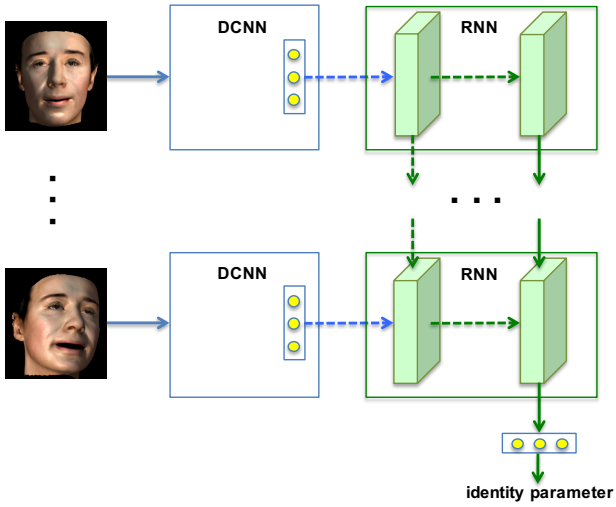


Figure 4: The depiction of the proposed RNN that consists of two recurrent layers and a linear fully-connected layer. The input to the RNN is the feature vector $\boldsymbol{g}_d$ from the DCNN. The output is the facial identity parameter vector.

3.3. Network pre-training and fine-tuning

To train the deep recurrent neural network, we follow the recent work [20, 21, 4] to generate synthetic 3D faces and 2D images. Similar to [4], we first create neutral 3D faces using random facial identity parameters. Then, for each neutral 3D face, we generate different facial expressions

by randomly sampling facial expression parameters from a pool of parameters estimated from real 2D face databases and synthesize 2D facial images with random illumination and facial pose. Following [4], we first transfer the weights of the convolutional layers from the VGG-Face [16] model trained on millions of real 2D images to our neural network. Then, we fine-tune the DCNN for single-view 3D face reconstruction. When the DCNN gets stabilized, we freeze all the convolutional layers and start to train the RNN and the fully-connected layers of the DCNN jointly for multi-view 3D face reconstruction. For the training loss, we choose to use the difference between the predicted 3D face and the ground truth. It is measured as the sum of squared error over all vertices of the 3D faces corresponding to the set of images:

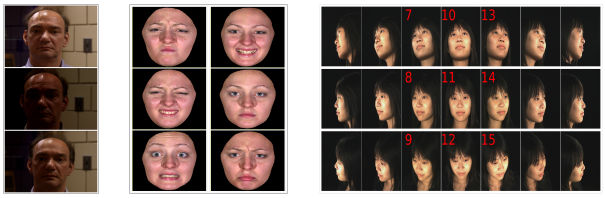
$$E = \sum_{i=1}^{N} (\|\boldsymbol{U}_d \cdot (\hat{\boldsymbol{\alpha}}_d - \boldsymbol{\alpha}_d)\|_2^2 + \lambda_e \|\boldsymbol{U}_e \cdot (\hat{\boldsymbol{\alpha}}_e^i - \boldsymbol{\alpha}_e^i)\|_2^2) \ , \tag{2}$$

where N is the number of facial images in a single set, $\hat{\boldsymbol{\alpha}}_d$ and $\hat{\boldsymbol{\alpha}}_e^i$ denote the predicted parameter vectors, $\boldsymbol{\alpha}_d$ and $\boldsymbol{\alpha}_e^i$ denote the ground truth, and λ_e denotes the factor controlling the importance of facial expression reconstruction.

4. Experiments

Databases: In this section, we evaluate our *DRFAR* algorithm for multi-view 3D face reconstruction. We use three publicly available 3D face databases in our experiments, namely the FRGC2 database [18], the BU-3DFE database [32], and the UHDB31 database [30]. For the FRGC2 database, we use the validation partition that consists of 4,007 pairs of 2D and 3D data of 466 subjects. The 2D facial images are captured under different illumination conditions. For the BU-3DFE database, we use all 2,500 pairs of 2D and 3D data of 100 subjects. The 2D and 3D data are captured while the subjects are performing different types of facial expressions. For the UHDB31 database, we use all the 4,851 2D facial images with corresponding 3D facial scans. These data are captured under three illumination conditions with 21 facial poses. We show several examples of the 2D data from these three databases in Fig. 5.

Metrics: We used the root mean squared error between the reconstructed 3D face and the ground truth after rigid alignment and registration using the iterative closest point (ICP) algorithm [1] implemented in MATLAB to measure the quantitative accuracy of 3D face reconstruction. We also use visual results of reconstructed 3D faces to qualitatively compare the reconstruction accuracy of different algorithms. To measure the consistency of facial identity reconstruction, we follow [27] to perform face recognition with the estimated facial identity parameters and use the rank-1 identification rate as a measurement.



(a) (b) (c)

Figure 5: Example of 2D images from the three public databases used in experiments: (a) The FRGC2 database, (b) the BU-3DFE database, and (c) the UHDB31 database.

Implementation details: We use the Dlib DNN-based face detector[1] for face detection and use the Caffe framework [9] to train and deploy the neural networks. The Adam solver [13] is employed with the mini-batch size and the initial learning rate set to 36 and 0.0001, respectively. Within each mini-batch, there are six different subjects, each with six facial images. Although *DRFAR* is trained for 3D face reconstruction with six images, it can be deployed for face reconstruction with a different number of images. When the number of facial images is less than six, they will be duplicated and fed into the neural network repeatedly. When the number of facial images is more than six, all of them will be fed into the neural network. In our experiment, for simplicity, we test *DRFAR* for 3D face reconstruction with $N = \{1, 3, 6\}$ images. The weighting factor is set to $\lambda_e = 5$. The training of the DCNN runs for 160,000 iterations while the training of the RNN runs for 32,000 iterations. We compare *DRFAR* with *E2FAR* [4], *RSNIEF* [20], *DRSN* [27], *2FCSL* [5], and *AFAR* [23]. Except for *DRFAR* and *E2FAR*, all other methods listed above require facial landmarks for 3D-2D alignment, and thus are not applicable for facial images with large pose because landmark detection will fail. We use the Dlib facial landmark detector for these methods.

4.1. Single-view 3D face reconstruction

In the first experiment, we apply *DRFAR* for 3D face reconstruction by using only a single image. The goal is to evaluate the generalization capability of our approach for both single-view and multi-view 3D face reconstruction. To this end, we use the UHDB31 database, and run *DRFAR* and *E2FAR* on all the 21 facial poses. For *RSNIEF*, *DRSN*, and *2FCSL*, however, we test on the subset UHDB31B, which consists of only the central nine facial poses as illustrated in Fig. 5(c). We do not use the large pose facial images because facial landmark detection fails for most of them. The quantitative results of mean and standard deviation of

[1] http://dlib.net/

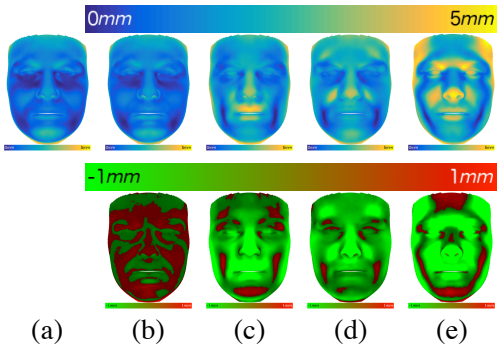


Figure 6: Single-view 3D face reconstruction error heatmaps of different methods on the UHDB31B database: (a) *DRFAR*, (b) *E2FAR* [4], (c) *RSNIEF* [20], (d) *2FCSL* [5], and (e) *DRSN* [27]. The top row illustrates the spatial distribution of RMSE on the face and the bottom row illustrates the differences in RMSE between our approach *DRFAR* and other approaches (green indicating our method has smaller RMSE, red indicating our method has larger RMSE, and color intensity indicating the magnitude of the difference).

RMSE are illustrated in Table 1. The spatial distributions of per-vertex reconstruction error of the four methods over the facial region are depicted in Fig. 6 (T) and the comparisons between our method *DRFAR* and the other methods are depicted in Fig. 6 (B). Compared with *E2FAR*, our method has slightly larger mean RMSE. In the inner facial region including the mouth and eyes, however, our method exhibits better performance than *E2FAR*. Compared with the other methods, both *DRFAR* and *E2FAR* achieve significant improvement in reconstruction accuracy.

Table 1: Quantitative comparison of *DRFAR* and state-of-the-art algorithms on single-view 3D face reconstruction with the UHDB31 database: Mean and standard deviation of RMSE (*mm*).

	DRFAR	*E2FAR* [4]	*RSNIEF* [20]	*DRSN* [27]	*UH-2FCSL* [5]
UHDB31B	2.80±0.69	**2.73±0.71**	3.51±0.84	3.65±0.91	3.37±0.76
UHDB31	2.98±0.72	**2.89±0.74**	N/A	N/A	N/A

4.2. Multi-view 3D face reconstruction

In the second experiment, we apply *DRFAR* for 3D face reconstruction by using a set of $N = \{3, 6\}$ facial images. The goal is to evaluate DRFAR's performance for multi-view 3D face reconstruction from a small image collection. To this end, we use the UHDB31, the FRGC2, and the BU-3DFE databases. For each database, we create randomly a list of image sets, each with $N \in \{3, 6\}$ facial images of the same subject. For each image set, we run *DRFAR* to estimate a single facial identity parameter vector and N facial expression parameter vectors. Then, the 3D faces are reconstructed via Eq. 1. For *E2FAR*, we run it on each single image to estimate both the facial identity and the facial expression parameter vectors. The quantitative results of mean and standard deviation of RMSE are illustrated in Table 2. Compared with *E2FAR*, our method consistently improves the reconstruction accuracy on all three databases. The spatial reconstruction error distributions and the comparison between both methods are illustrated in Fig. 8. Compared with *E2FAR*, our method demonstrates better performance over large facial regions including several key facial components, such as the mouth, nose, and eyes. To analyze how our approach improves the reconstruction

Table 2: Quantitative comparison of multi-view 3D face reconstruction error on the UHDB31B database: Mean and standard deviation of RMSE (*mm*). Each set consists of $N = \{3, 5\}$ facial images of the same identity.

	N=3		N=6	
	DRFAR	*E2FAR* [4]	*DRFAR*	*E2FAR* [4]
UHDB31	**2.87±0.68**	2.98±0.72	**2.81±0.66**	2.92±0.74
FRGC2	**3.43±2.09**	3.46±2.11	**3.40±2.07**	3.48±2.14
BU-3DFE	**4.37±1.03**	4.48±1.09	**4.39±1.01**	4.48±1.10

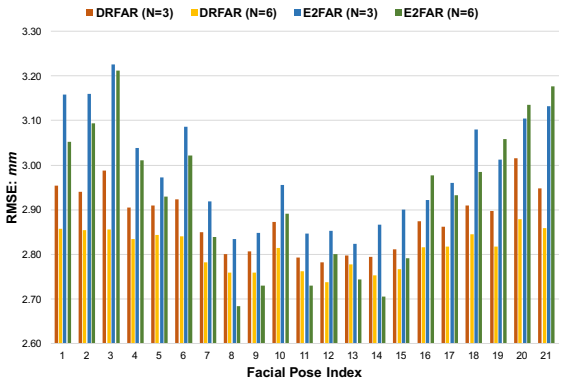


Figure 7: Fine-grained results of reconstruction RMSE on each specific facial pose category of the UHDB31 database.

Table 3: The rank-1 face identification rates (%) on the UHDB31 database.

	N=3			N=6		
	DRFAR	*E2FAR* [4]	*DRSN* [27]	*DRFAR*	*E2FAR* [4]	*DRSN* [27]
split 1	**61.92**	53.67	49.00	**77.93**	62.61	52.70
split 2	**64.14**	56.57	50.11	**81.53**	60.81	54.95
split 3	**60.13**	57.24	48.55	**79.28**	61.71	54.50
split 4	59.69	**60.58**	48.11	**79.28**	61.71	54.05
split 5	**63.70**	59.02	49.44	**77.03**	63.06	55.41

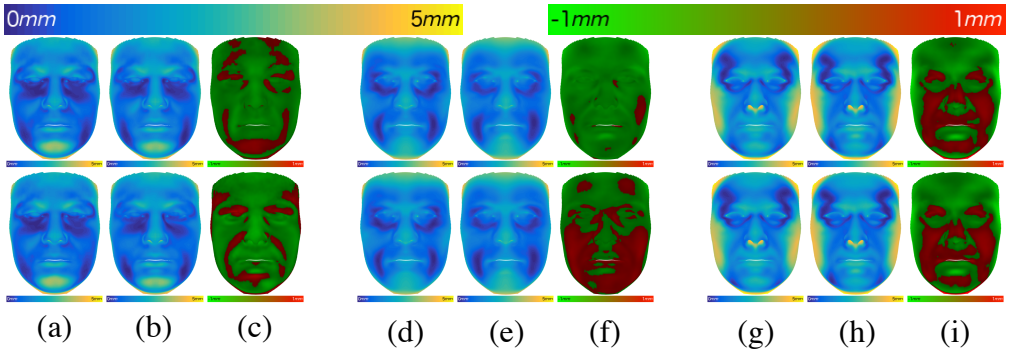


Figure 8: Reconstruction error heatmaps of *DRFAR* and *E2FAR* on three databases: UHDB31, FRGC2, and BU-3DFE. The top row illustrates the error heatmaps of both methods with $N = 3$ and the bottom row corresponds to $N = 6$. For each column: (a,d,g) depict the error heatmaps of *DRFAR* on the three databases, (b,e,h) depict the error heatmaps of *E2FAR* on the three databases, and (c,f,i) depict the differences in reconstruction error between our approach *DRFAR* and *E2FAR* (green indicating our method has smaller reconstruction error, red indicating our method has larger reconstruction error, and color intensity indicating the magnitude of the difference).

accuracy for large facial pose images, we also illustrate in Fig. 7 the mean RMSE for each specific facial pose category in the UHDB31 database. We can observe that our approach exhibits much lower RMSE for large pose facial images. For near-frontal facial images with very small pose variation, *E2FAR* performs better. We can also observe that, compared with using $N = 3$ images in a set, the reconstruction RMSE is further reduced by using $N = 6$ images in a set, which indicates that our approach can effectively aggregate the identity specific contextual information.

4.3. Face recognition based on facial identity parameters

Although accuracy is an important indicator for the performance of 3D face reconstruction, it is not sufficient. Another important indicator is the consistency or intra-similarity of the reconstructed 3D faces. In other words, reconstructed 3D faces of the same subject should be similar to each other. Following [27], in the third experiment, we perform face recognition on the UHDB31 database by using the estimated facial identity parameters. For each subject, out of the 21 facial poses, we use only the frontal facial image as gallery. The remaining non-frontal facial images are split randomly into multiple sets of $N = \{3, 6\}$ images and we use each image set as probe. For *DRFAR*, we use the single facial identity parameter vector estimated for each image set. For *E2FAR* and *DRSN*, we average the facial identity parameter vectors estimated for each image in the set. We repeat the experiment five times, each with a random split of the probe image set. The rank-1 identification rates are illustrated in Table 3. Compared with *E2FAR* and *DRSN*, our approach achieves significant improvement in face identification accuracy. The result indicates that, compared with averaging over a set of facial identity parameters estimated from each single image independently, our approach greatly improves the consistency of facial identity parameter estimation and the intra-similarity of 3D face reconstruction for the same subject.

4.4. Face reconstruction in-the-wild

In this experiment, we use facial images collected from the Internet for 3D face reconstruction. We select four celebrities and download, for each of them, six facial images with large pose variation using the Google image search engine. We compare *DRFAR* with *E2FAR* and a state-of-the-art algorithm, *AFAR* [23], for multi-view 3D face reconstruction. Compared with *DRFAR* and *E2FAR*, *AFAR* can only reconstruct a single 3D face for each image set. Examples of the reconstructed 3D face are illustrated in Fig. 9. Note that, for each set of facial images, the frontal facial image is used only for visualization and not used during 3D face reconstruction. The single 3D face beside each frontal facial image is reconstructed by *AFAR*. We can observe that *AFAR* performs poorly in this task. This is due to the small number of facial images available and the challenging pose variation exhibited. The first and second rows below the facial images depict the 3D faces reconstructed by *DRFAR* and *E2FAR*, respectively. Compared with *E2FAR*, as highlighted in red rectangles the inaccurate reconstruction, the performance of our method is more robust against adverse factors in image quality and more consistent in the reconstructed 3D facial shape.

5. Conclusions

In this paper, we explore the problem of multi-view 3D face reconstruction and propose an algorithm based on deep recurrent neural networks to learn a nonlinear mapping

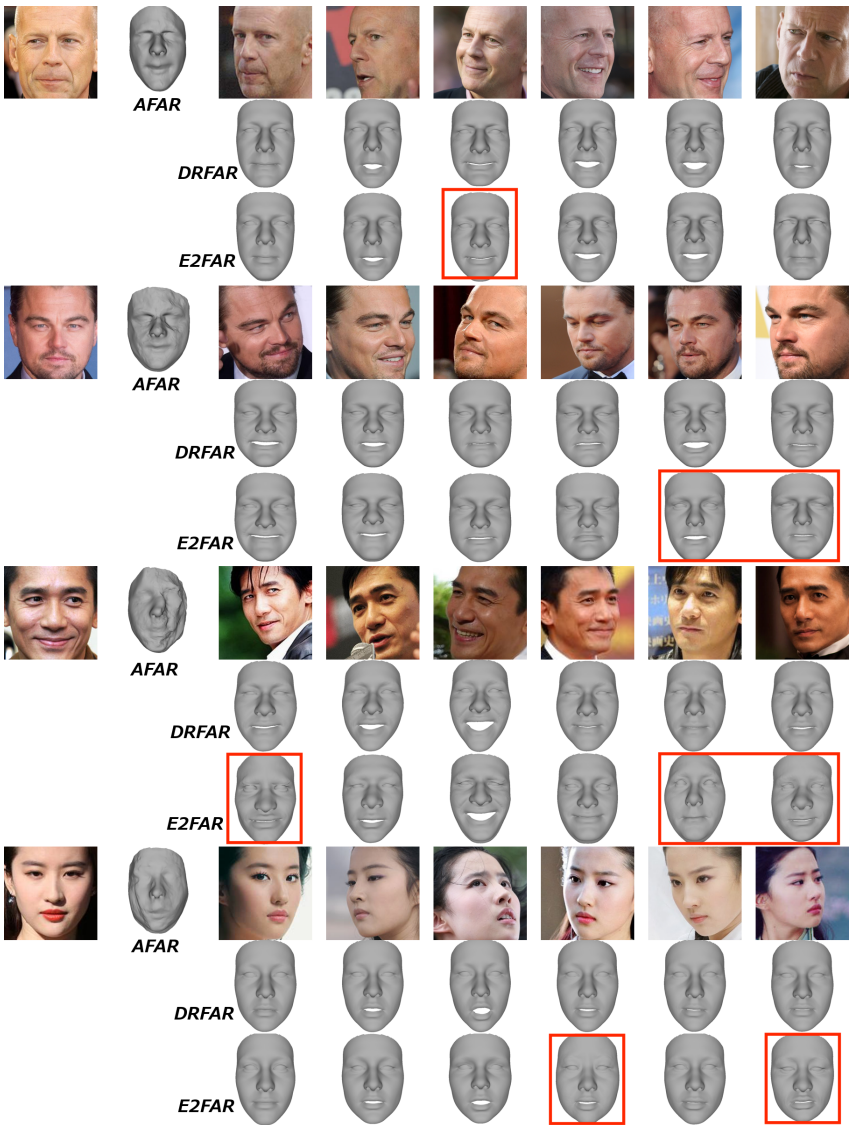


Figure 9: Example of the 3D faces reconstructed by different methods: *DRFAR*, *E2FAR* [4], and *AFAR* [23] (Please refer to the text for more details).

from a small set of facial images captured under different conditions to a subspace representation of the corresponding 3D faces. In our method, we treat the identity signal of the subject in images as contextual information and propose using LSTM cells in an RNN to extract the identity specific contextual information by fusing and aggregating the identity-related features extracted from each facial image by a DCNN. Through extensive experiments, we demonstrate the superiority of our method in improving both the accuracy and the consistency of 3D face reconstruction using a very small set of $N = \{3, 6\}$ facial images.

Acknowledgment

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH001. This grant is awarded to the Borders, Trade, and Immigration (BTI) Institute: A DHS Center of Excellence led by the University of Houston, and includes support for the project "Image and Video Person Identification in an Operational Environment: Phase I" awarded to the University of Houston. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

[1] P. Besl and N. McKay. A method for registration of 3D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, February 1992.

[2] A. Brunton, J. Lang, E. Dubois, and C. Shu. Wavelet model-based stereo for fast, robust face reconstruction. In *Proc. Canadian Conference on Computer and Robot Vision*, pages 347–354, St. Johns, NL, May 25-27 2011.

[3] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 677 – 691, Boston, MA, June 2015.

[4] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, Honolulu, Hawaii, July 22-25 2017.

[5] P. Dou, Y. Wu, S. K. Shah, and I. A. Kakadiaris. Robust 3D facial shape reconstruction from single images via two-fold coupled structure learning. In *Proc. British Machine Vision Conference*, pages 1–13, Nottingham, United Kingdom, September 1-5 2014.

[6] Y. Gao and D. Glowacka. Deep gate recurrent neural network. In *Proc. Asian Conference on Machine Learning*, page 350365, The University of Waikato, Hamilton, November 16-18 2016.

[7] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[8] A. Ichim, S. Bouaziz, and M. Pauly. Dynamic 3D avatar creation from hand-held video input. *ACM Transactions on Graphics*, 45:1–14, 2015.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. International Conference on Multimedia*, pages 675–678, Orlando, Florida, USA, Nov. 03 - 07 2014.

[10] A. Jourabloo and X. Liu. Large-pose face alignment via CNN-based dense 3D model fitting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4188 – 4196, Las Vegas, NV, June 26-July 1 2016.

[11] I. Kemelmacher-Shlizerman. Internet-based morphable model. In *Proc. IEEE International Conference on Computer Vision*, pages 3256–3263, Sydney, Australia, December 1-8 2013.

[12] I. Kemelmacher-Shlizerman and S. Seitz. Face reconstruction in the wild. In *Proc. IEEE International Conference on Computer Vision*, pages 1746–1753, Barcelona, Spain, Nov. 6-13 2011.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations*, pages 1–15, San Diego, CA, May 7-9 2015.

[14] S. Lee, K. Park, and J. Kim. A SfM-based 3D face reconstruction method robust to self-occlusion by using a shape conversion matrix. *Pattern Recognition*, 44(7):1470 – 1486, July 2011.

[15] Y. Lin, G. Medioni, and J. Choi. Accurate 3D face reconstruction from weakly calibrated wide baseline images with profile contours. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1490–1497, San Francisco, CA, June 13-18 2010.

[16] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. British Machine Vision Conference*, pages 1–12, Swansea, UK, September 7-10 2015.

[17] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *Proc. International Conference on Machine Learning*, pages 1–12, Atlanta, GA, June 16-21 2013.

[18] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):831–846, May 2010.

[19] R. Rana. Gated recurrent unit (GRU) for emotion classification from noisy speech. *ArXiv e-prints*, pages 1–9, December 2016.

[20] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *Proc. International Conference on 3D Vision*, pages 460–469, California, USA, October 25-28 2016.

[21] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, Honolulu, Hawaii, July 22-25 2017.

[22] J. Roth, Y. Tong, and X. Liu. Unconstrained 3D face reconstruction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2606–2615, Boston, USA, June 7-12 2015.

[23] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, Las Vegas, NV, June 26-July 1 2016.

[24] B. Shi, K. Inose, Y. Matsushita, P. Tan, S. Yeung, and K. Ikeuchi. Photometric stereo using internet images. In *Proc. International Conference on 3D Vision*, pages 361–368, Tokyo, Dec. 8-11 2014.

[25] B. Shi, Y. Matsushita, Y. Wei, C. Xu, and P. Tan. Self-calibrating photometric stereo. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1118–1125, San Francisco, CA, June 13-18 2010.

[26] P. Snape, Y. Panagakis, and S. Zafeiriou. Automatic construction of robust spherical harmonic subspaces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 91–100, Boston, Massachusetts, June 7-12 2015.

[27] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, Honolulu, Hawaii, July 22-25 2017.

[28] C. Wu. Towards linear-time incremental structure from motion. In *Proc. International Conference on 3D Vision*, pages 127–134, Seattle, WA, June 2013.

[29] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Proc. Asian Conference on Computer Vision*, volume 6494, pages 703–717, Queenstown, New Zealand, November 8-12 2011.

[30] Y. Wu, S. K. Shah, and I. A. Kakadiaris. Rendering or normalization? An analysis of the 3D-aided pose-invariant face recognition. In *Proc. IEEE International Conference on Identity, Security and Behavior Analysis*, pages 1–8, Sendai, Japan, Feb. 29-Mar. 2 2016.

[31] C. Yang, J. Chen, C. Xia, J. Liu, and G. Su. A SfM-based sparse to dense 3D face reconstruction method robust to feature tracking errors. In *Proc. IEEE International Conference on Image Processing*, pages 3617–3621, Melbourne, Australia, September 15-18 2013.

[32] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3D facial expression database for facial behavior research. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 211–216, Southampton, UK, Apr. 10-12 2006.

[33] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 146 – 155, Las Vegas, NV, June 26-July 1 2016.