# Improving 2D Face Recognition via Discriminative Face Depth Estimation

Jiyun Cui[1,2], Hao Zhang[1,2], Hu Han[*,1], Shiguang Shan[1,2,3], and Xilin Chen[1,2]

[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
{jiyun.cui, hao.zhang}@vipl.ict.ac.cn {hanhu, sgshan, xlchen}@ict.ac.cn

## Abstract

*As face recognition progresses from constrained scenarios to unconstrained scenarios, new challenges such as large pose, bad illumination, and partial occlusion, are encountered. While 3D or multi-modality RGB-D sensors are helpful for face recognition systems to achieve robustness against these challenges, the requirement of new sensors limits their application scenarios. In our paper, we propose a discriminative face depth estimation approach to improve 2D face recognition accuracies under unconstrained scenarios. Our discriminative depth estimation method uses a cascaded FCN and CNN architecture, in which FCN aims at recovering the depth from an RGB image, and CNN retains the separability of individual subjects. The estimated depth information is then used as a complementary modality to RGB for face recognition tasks. Experiments on two public datasets and a dataset we collect show that the proposed face recognition method using RGB and estimated depth information can achieve better accuracy than using RGB modality alone.*

## 1. INTRODUCTION

Face recognition has achieved great progress in the past decades of years and is becoming usable in many real application scenarios, such as checking-in systems, security departments, and law enforcement. While most of the current face recognition methods are focusing on 2D images, RGB-D or 3D based face recognition shows more robustness against large pose, partial occlusion, and uneven illumination variations [5, 6, 3, 9]. However, even the gallery images of the face recognition systems can be captured with RGB-D sensors to retain more subject discriminative information, it is not possible to replace all the existing RGB cameras with RGB-D sensors in a short time.

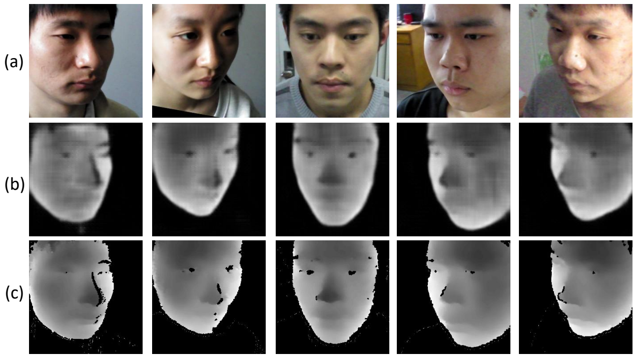

Figure 1. Discriminative face depth estimation from 2D RGB face images. (a) input RGB images, (b) estimated depth image by our approach, and (c) the ground-truth depth. The depth images are scaled into the range of [0, 255].

In addition, a large amount of 2D face databases will be continuously used by existing face recognition systems for a long time. In this situation, a more feasible way is that the enrollment of individual subjects in the gallery set uses RGB-D, but the testing images remain in RGB modality. This means that the existing face recognition systems need to handle both RGB-D multi-modality gallery images and RGB unimodality probe images. This is a challenging task, and there is only a limited number of studies on this topic.

To match RGB probe images with RGB-D gallery images, there are three possible solutions: (i) directly throw the depth modality of the gallery face images, and perform RGB probe vs. RGB gallery matching; (ii) use canonical correlation analysis or common feature learning methods to conduct RGB probe vs. depth gallery problem, and then fuse the model with RGB probe vs. RGB gallery model; (iii) use a depth estimation model to estimate the needed

[*] H. Han is the corresponding author.

depth images for the probe set, and then apply RGB-D probe vs. RGB-D gallery algorithm. Among these approaches, the first approach does not exploit the complementary information provided by depth modality. The second approach faces great challenges because the RGB and depth are very different modalities. The third approach has much wider application scenarios, e.g., by converting RGB vs. RGB-D face recognition into RGB-D vs. RGB-D face recognition (see Fig. 1). The original modality gap between 2D probe images and RGB-D gallery images no longer exist, and thus, all the existing RGB-D face recognition algorithms can be used. In this paper, we focus on the study of depth estimation from 2D face images to solve the RGB vs. RGB-D face matching problem.

However, depth estimation from a single RGB face image is nontrivial. The main difficulties of this problem are: (i) Similar to 3D reconstruction from a single 2D image, depth image estimation from a single RGB image is also an ill-posed problem. This is because an object with different 3D shape may still generate the same 2D image. (ii) RGB and depth modalities are heterogeneous from each other. While RGB image represents the texture information, the depth reflects the shape information. (iii) The estimated depth is expected to be not only visually reasonable but also discriminative for face recognition tasks.

Inspired by the pixel-wise labeling ability of fully convolutional network (FCN) [24] and discriminative feature learning ability of convolutional neural network (CNN) [25, 28, 10], we propose a network architecture of cascaded FCN and CNN for discriminative depth estimation approach from 2D face images. Specifically, FCN aims to predict the depth information from an RGB image, while CNN aims to retain the subject discriminative information in the estimated depth. In the cascaded network framework, the feature map of the last layer in FCN is fed into the first layer in CNN, which connects FCN and CNN and as the face depth information estimated. There are two loss functions in the proposed depth estimation network: Euclidean loss penalizing the errors between the estimated depth and the ground truth depth, and a softmax loss [16] enforcing the discriminative ability of the estimated depth information. Both the Euclidean loss and the softmax loss are back propagated to the whole cascaded network. An overview of the proposed approach is shown in Figure 2. Experimental results on the public domain Lock3DFace [31] and IIIT-D RGB-D face database [5], and a database we collected show that the proposed face depth estimation is very effective in improving the face recognition accuracy than single RGB modality.

The main contributions of this work are: (i) a cascaded FCN and CNN architecture for discriminative depth estimation from a single RGB images; (ii) the ability of mitigating the modality gap for matching RGB probe face images with RGB-D gallery face images; and (iii) extending the application scope for current RGB-D face recognition methods into cross-modality scenarios.

## 2. RELATED WORK

### 2.1. Face Depth Estimation

The constrained independent component analysis (cICA) model was proposed in [26] to convert the overcomplete ICA problem into a normal ICA problem for face depth estimation from one or various different posed 2-D images. [27] used a nonlinear least-squares model with similarity transform for the pose of the face image and different optimization schemes to predict the 3D structure of the human face. The methods proposed in [26, 27] utilized various pose of 2D face images to depth estimation. [32] can estimate face depth from the single frontal face image. The method proposed in [32] takes face depth estimation as a statistical learning problem. To find the transformation of face texture image and depth image, the authors used the local binary pattern (LBP) model to encode the face texture. The statistical method proposed by [23] is based on Canonical Correlation Analysis (CCA) and estimate depth information from the frontal color face image, but the method has a limitation on the angle of the color face. A framework was proposed in [15] for depth estimation, this framework consists a 2D-3D database which formed by feature points of 2D and 3D images. Based on the 2D-3D database, the method proposed in [15] can estimate depth from single test 2D image.

Different from the above methods, the proposed approach for face depth estimation from a single color image is based on a cascaded FCN and CNN architecture. Our face depth estimation method aims at not only minimizing the differences between the estimated depth and the ground-truth depth but also retaining as much subject discriminative information as possible.

### 2.2. General Image Depth Estimation

In addition to face depth estimation, there are also a number of methods for depth estimation from a generic image. In [30], the scenario structure was used to estimate the absolute mean depth of the scene. Similar to our approach, the estimated depth image is used to recognition. However, our method aims to predict a depth image that retains the subject-discriminative information. A convolutional neural network was designed in [2] to predict the relative depth of a single image in the wild. An FCN with residual learning is proposed in [17] to predict the depth information in an end-to-end manner. The method in [1] also used a deep fully convolutional residual networks for depth estimation. But the difference is that [1] treat depth estimation as a classification problem instead of regression problem and take

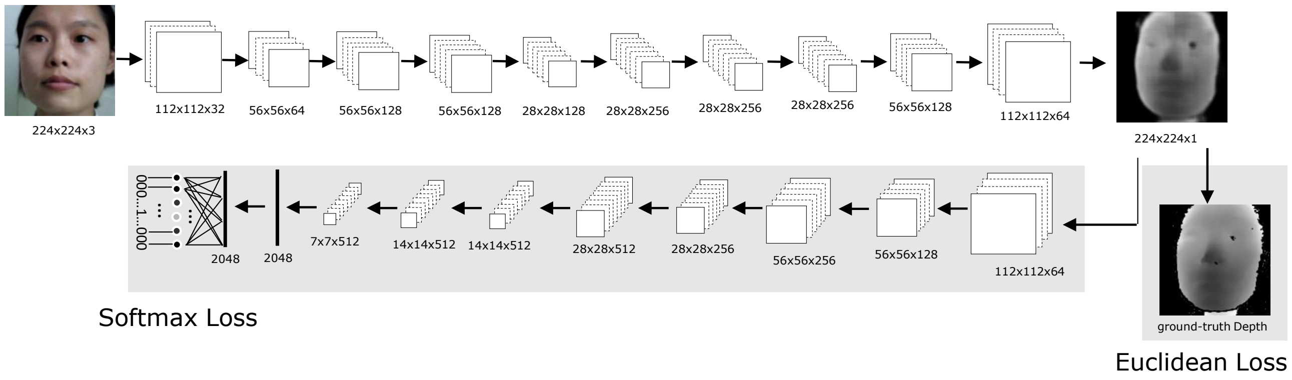

Figure 2. An overview of the proposed approach for discriminative face depth estimation.

post-processing to improve the performance. [12] proposed a method to computer projection error, which can improve the performance of depth estimation by reinforcing the gradient component. Conditional Random Field network was exploited in [19, 20, 11, 18] for depth estimation of general scenarios. The methods proposed in [19] and [20] combined the capacity of deep CNN and continuous CRF, which construct a deep CNN framework to learning the potential of continuous CRF, aim to estimate depth information from a single monocular image. Convolutional Conditional Random Field Network (CCRFN) model was proposed in [11], which used two CNN to learn the absolute and relative features from raw images and then fed learned features to CRF to generate the depth of images. A deep convolutional neural network (DCNN) was proposed in [18] to regress the depth and surface normal information of an image and use CRF to refine the result to improve the performance. A coarse-to-fine CNN was proposed in [4] which used a coarse-scale network to predict the global depth information and a fine-scale network to refine the coarse depth map.

We can see that CNN based approaches have been found to be effective for general scenario depth estimation from a single 2D image. In this paper, we also use CNN for face depth estimation from a single face image. But there are a few differences. Firstly, we use CNN to predict the depth of a particular object, rather than the general scene. Secondly, we aim to produce more distinguishable depth information of face image instead of less different with ground-truth depth image. Thirdly, we combine estimated depth face image and color face image to face recognition, the experiment shows that it is better than using single color modality.

### 2.3. RGB-D Face Recognition

As the depth cameras become more and more popular, a few RGB-D face image database is available public, there are some methods to exploit the RGB-D face recognition. Related to RGB-D face recognition, [13] proposed the method which using RGB-D face data to gender recognition based on the LBP descriptor of gray and depth image. [5] compute the RGB entropy and depth entropy of face image as well as the HOG of visual saliency map of the human face, the random decision forest as the classifier. [6] is based on the methods proposed by [5], apart from the descriptor based on the entropy of RGB-D image and saliency map, they also extract the geometric facial attribute from depth face image, the result of descriptor and attribute is fused for RGB-D face recognition. The most related work to our approach is [3], which used an autoencoder to estimate the depth information. The proposed approach differs from [3] in that we are trying to not only minimize the depth estimation errors but also retain the discriminative ability of the estimated depth.

## 3. PROPOSED APPROACH

### 3.1. Overview

Based on the observation that most of the published depth estimation methods for 2D images aim to minimize the Euclidean loss ($L_2$ loss) between the estimated depth and the ground-truth depth. Such a loss is able to retain the subject's identity information to some extent but still lacks explicit constraint to assure high discriminative ability among individual subjects.

We choose to jointly consider the depth reconstruction error in terms of $L_2$ loss and subject classification error in terms of softmax loss. Formally, the goal of our approach is to minimize

$$\min_{W_F,W_C} |d-\mathrm{F}_F(x,W_F)|_2+log\sum_j e^{z_j}-z_y \tag{1}$$

$$\mathbf{z}=\mathrm{F}_C(\mathrm{F}_F(x,W_F),W_C) \tag{2}$$

where $x$ is the input 2D face image and $d$ is the ground-truth depth face image. $\mathrm{F}_F$ and $\mathrm{F}_C$ denote the depth estimation function and multi-class classification function, respectively. $y$ is the label corresponding to the input data $x$. $W_F$

and $W_C$ are the parameters for the depth estimation function and multi-class classification function, respectively. For the depth estimation function $\mathrm{F}_F$, we choose to use an FCN network, and for the multi-class classification function $\mathrm{F}_C$ we choose to use CNN network. With the joint loss by FCN and CNN, the depth estimation process is forced to replicate the ground-truth depth information while retaining subject discriminative information. An overview of the proposed approach is shown in Figure 2.

### 3.2. Implementation Details

**Depth estimation with FCN**. FCN was first proposed as a classification network for image semantic segmentation by pixel-wise labeling, e.g., foreground vs. background. In this paper, FCN is utilized for a more fine-grained pixel-wise labeling, i.e., assign a depth value from [0, 255] to each pixel. As shown in Figure 2, our FCN network consists of eight convolutional layers and three deconvolutional layers. We use $2 \times 2/s2$ max pooling to reduce the feature map size to $1/4$ times pixels of previous layer (s2 means stride is set to 2), and use $6 \times 6/s2$ deconvolution layer to increase the feature map size to 4 times pixels of the previous layer. Besides, all convolution layers do not change the feature map size, and they are all $3 \times 3/s1$ kernels. We use PReLU as the non-linear rectifier function after each convolution layer and deconvolution layer, except the last deconvolution layer. BatchNormalization (BN) [14] layer is added before every rectifier. The input RGB image to FCN is $224 \times 224 \times 3$, the target depth image is $224 \times 224 \times 1$ which is a one-channel depth face image. Euclidean loss is used to penalize the error between the estimated depth and the ground-truth depth.

**Subject discriminative learning via CNN**. To enforce the estimated depth image to retain as much subject discriminative information as possible, we propose to use a CNN to penalize the inter-subject similarities. Specifically, we utilize a VGG-11 network [25], which has eight convolutional layers, three fully connected (FC) layers, and with a BN layer added before each ReLU rectifier in the network. In addition, we change the two 4096D FC layers in VGG-11 into 2,048 FC layers.

The input of VGG-11 is the output of the preceding FCN network which is a one channel depth image, the output of CNN network is the probability of every face label. The feature maps size of every layer is annotated below corresponding network layers shown in Figure 2.

**Network training**. The network weights are updated with the following policy

$$\Delta_i W = m\Delta_{i-1} W + \omega_1 \frac{\partial L_F}{\partial W_F} + \omega_2 \frac{\partial L_C}{\partial W} \tag{3}$$

where $\Delta_i W$ is the transformation weights update tensor of $i$-th iteration, $\omega_1$ and $\omega_2$ is the weights of loss $L_F$ and $L_C$. Generally, $\omega = \omega_{loss} \cdot \alpha$, $\omega_{loss}$ is the loss weight and $\alpha$ is learning rate. $m$ is momentum parameter.

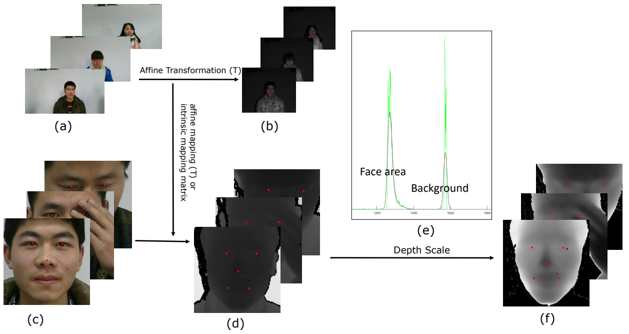
(a)
Affine Transformation (T)
(b)
affine mapping (T) or intrinsic mapping matrix
Face area
Background
1200
1400
1600
1800
(e)
Depth Scale
(c)
(d)
(f)

Figure 3. The processing for the RGB and depth image: (a) the original RGB images and the detected face landmarks, (b) the corresponding near infrared images (the same size as the depth images) and face landmarks, (c) cropped RGB face images, (d) cropped depth face image based on mapping matrix, (e) the histogram of cropped and smoothed depth image, (f) scale depth image into [0, 255].

We notice that the scale of the $L_2$ loss in FCN is much larger than the scale of the CNN softmax loss. We set the loss weight of FCN to 0.001 to balance the influence of $L_2$ loss and softmax loss. Then, our cascaded FCN and CNN network can be trained in an end-to-end approach.

**Face Recognition**. Given the RGB face images and the estimated depth images, we train an Inception V2 network [14] for each modality and extract the 1024D features from the loss1/fc layer for the Inception V2 network. We then explore three types of modality fusion methods: score-level fusion, feature-level fusion, and channel-level fusion. In score-level fusion, we calculate the correlation similarity of each modality, and fuse the RGB and depth scores using a sum rule

$$S = \alpha \cdot S_1 + (1 - \alpha) \cdot S_2 \tag{4}$$

where $\alpha$ is the weight for the RGB modality, and we use $\alpha = 0.5$ by default. In feature-level fusion, we concatenate the RGB and depth features together to get a 2048D feature vector, and use the correlation of two features as the similarity. In the channel-level fusion, depth is concatenated with RGB as a four-channel input data into the Inception V2 [14] network. The features extracted from the loss1/fc layer are the final features of a RGB-D image and correlation is used as the similarity.

**The preprocessing of RGB-D images**. Specifically, as shown in Figure 3 (a), we detect facial landmarks on RGB images using an open-source face recognition engine[1]. We can not perform landmark detection on the depth images because most of the face alignment algorithms are developed for RGB images. Instead, we map the landmark positions

[1] `https://github.com/seetaface/SeetaFaceEngine`.

from RGB to depth based on either the mapping matrix provided by the RGB-D sensors or the affine transformation matrix computed based on pairs of RGB and near-infrared images (see Figure 3 (b)) due to the near-infrared images and depth images share the same coordinate. The RGB and depth images are then cropped based on the five facial landmarks. Finally, we scale the cropped depth image into [0, 255].[2] For the Lock3DFace RGB-D database, there are a lot of noises in the depth image, so we used a bilateral filter [29] to suppress the noises.

# 4. Experimental Results

## 4.1. Databases

We provide evaluations on two public-domain RGB-D databases including IIIT-D [5], and BUAA Lock3DFace database [31], as well as an RGB-D database we collected.

**IIIT-D**. IIIT-D contains 4,603 RGB-D images of 106 subjects, which were captured using Kinect I with normal illumination and a few pose and expression variations. We randomly select 72 subjects for training dataset and the remaining 34 subjects for test dataset.

**BUAA Lock3DFace**. BUAA Lock3DFace has much more subjects than IIIT-D and contains 5,711 RGB-D video sequences of 509 subjects, which were recorded using Kinect II with variations in pose, expression, occlusion, and time. We extract 33,780 RGB-D images from all the video sequences and randomly select 340 subjects (22,798 images) as the training dataset, the remaining 169 subjects (10,982 images) RGB-D images are used for testing. We also divide the testing dataset into five subsets following [31], which are frontal, pose, expression, occlusion and time subsets, respectively.

**Our database**. Our database was captured using RealSense II instead of Kinect, containing about 845K RGB-D images of 747 subjects with continuous pose variations and a few illumination changes. We randomly select 500 subjects (about 581,366 images) for training and the remaining 247 subjects for testing. After division, the training set contains 581,366 pairs of images and testing set contains 280,257 pairs of images.

For each subject used for testing in the BUAA Lock3DFace, IIIT-D, and our dataset, one frontal or near-frontal face image is used as the gallery, and the other images are used as the probe. Table 1 summarizes the testing protocols of the three databases. We should point out that the training set of each database is used for training both the depth estimation model and the face recognition model.

Table 1. Database divisions for the experiments on IIIT-D, BUAA Lock3DFace, and our dataset.

| Database | Variations | Train (#img./#sub.) | Test (#img./#sub.) |
|---|---|---|---|
| IIIT-D | Mixed | 3,632/72 | 971/34 |
| BUAA Lock3DFace | Neutral-fromtal | 4,068/340 | 2,016/169 |
| | Expression | 5,172/340 | 2,544/169 |
| | Pose | 4,040/340 | 2,012/169 |
| | Occlusion | 3,909/340 | 1,915/169 |
| | Time(Session2) | 5,614/117 | 2,495/52 |
| | Total | 22,798/340 | 10,982/169 |
| Ours | Mixed | 581,366/500 | 280,257/242 |

## 4.2. Evaluation metrics

**Depth Estimation**. We compute the pixel-wise Mean Absolution Error (MAE) between the estimated depth image and the ground-truth depth image, i.e., $MAE = \frac{1}{n}\sum_{i=1}^{n}|d_i - d_i^*|$, where $n$ is the total number of pixels of the depth image, $d_i$ and $d_i^*$ are the estimated depth value and the ground-truth depth value of the $i$-th pixel.

**Face Recognition**. We report the rank-1 identification accuracy for comparing the face recognition performance with and without using the estimated face depth information by our method. In order to fully evaluate the accuracy of face recognition of baseline and proposed methods, we report the result of the three types of multi-modalities human recognition.

## 4.3. Results

We first report the face depth estimation accuracy by the proposed approach and compare it with the baseline method, i.e., FCN [24]. Figure 4 shows that both the proposed approach and FCN generate visually pleasing depth images from the input RGB images. The pixel-wise depth estimation MAEs on the Lock3DFace, IIIT-D, and our dataset are reported in Table 3. These results may raise doubts about the effectiveness of the proposed discriminative depth estimation method. The reason is that MAE mainly corresponds to the subjective quality w.r.t. the ground-truth depth, but a lower MAE does not necessarily mean higher discriminative ability. This observation is verified by our following face recognition experiments using RGB and estimated depth images.

Since the purpose of depth estimation is to improve the 2D face recognition performance, we perform face identification experiments on Lock3DFace, IIIT-D, and our dataset using the RGB images and the estimated depth information to verify whether our discriminative depth estimation approach is helpful for improving 2D face recognition accuracy or not.

We assume that the both the RGB and depth modalities are available in the training and gallery sets, but only RGB modality is available in the probe set. So on the training dataset, three Inception V2 models can be trained for RGB,

[2] A detailed description of our scaling pipeline is provided at: `http://www.escience.cn/people/hhan/publication.html`

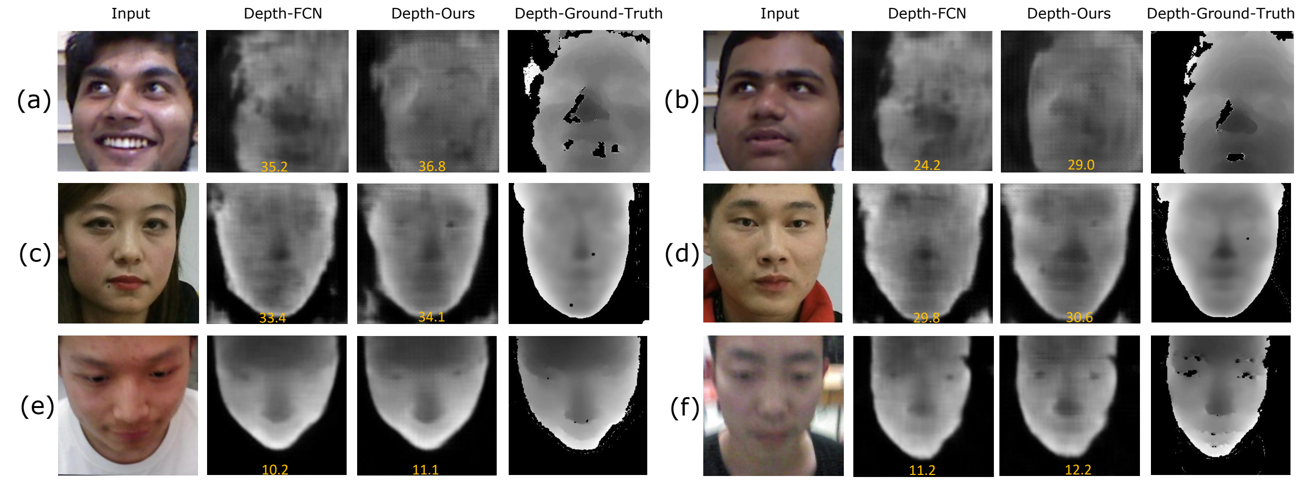

Figure 4. Example of the depth estimation results on (a-b) IIIT-D database, (c-d) the Lock3DFace database, and (e-f) our database. The four columns from left to right represent the RGB images, depth images estimated by FCN [24], depth images estimated by our method, respectively. The numbers under the estimated depth images represent the MAEs value.

Table 2. Definitions of the face recognition model type, gallery type, and probe type used in the face identification experiments.

| Model type | Description | Gallery type | Description | Probe type | Description |
|---|---|---|---|---|---|
| $M_1$ | Inception V2 FR model trained using ground-truth depth | $G_1$ | ground-truth depth | $P_1$ | ground-truth depth |
| $M_2$ | Inception V2 FR model trained using estimated depth | $G_2$ | estimated depth | $P_2$ | estimated depth |
| $M_3$ | Inception V2 FR model trained using ground-truth and estimated depth | - | - | - | - |

Table 3. Comparisons of the depth estimation errors by our approach and the baseline method on three RGB-D databases.

| Method | Pixel-wise MAE | | |
|---|---|---|---|
| | Lock3DFace | IIIT-D | Our Dataset |
| FCN [24] | 22.3 | 40.2 | 12.9 |
| Proposed approach | 22.4 | 40.1 | 13.4 |

ground-truth depth, estimated depth, respectively. For the gallery set, we can use either the ground-truth depth or the estimated depth. For the probe set, we focus on the evaluations of the estimated depth, and the ground-truth depth is only used as a reference for the *upper bound* of the performance. To simplify our descriptions, we provide explanations for the model, gallery, and probe types in Table 2.

The face identification performance using different combinations of model type and gallery type on Lock3DFace, IIIT-D, and our dataset are shown in Table 4. The face identification accuracies using RGB modality alone are 94.5%, 95.9%, and 93.3% on Lock3DFace, IIIT-D, and our dataset, respectively. When the estimated depth is used for face identification together with the RGB image, we can see the depth estimated by our method (M2/G2/P2) achieves the best result on the Lock3DFace database, and it improve the 2D face recognition accuracy from 94.5% to 94.90%. Similarly, the depth estimated by our approach leads to higher face recognition accuracy on IIIT-D (M2/G2/P2) and our dataset (M3/G1/P2), than the baseline depth estimation. We also report the face identification performance on the individual subsets of Lock3DFace in Table 5. The proposed approach shows excellent performance on all the subsets except for $Probe_S5$, which is a probe set with a time gap to the gallery set. The results by three different modality fusion methods (described in Sect. 3.2) are shown in Table 6. We can see that a score-level fusion of RGB and the estimated depth by our approach can consistently improve the face recognition performance.

We also perform a T-test between the accuracies of RGB face recognition and our multi-modality face recognition. The p-value is $0.1$, which suggests that the improvement by the proposed approach is significant. These results show the effectiveness of the our approach in retaining subject discriminative information in the estimated depth images, and improve the face recognition accuracy compared to using only the RGB face images. Theoretically, the depth images are estimated from the RGB images, and do not increase any information. The possible reason why depth estimation

Table 4. Face identification accuracies on the Lock3DFace, IIIT-D, and using the estimated depth information by the proposed approach and the baseline method. The results using RGB and the ground-truth depth images are shown in red, and used as the reference for the upper bound of the performance using RGB and estimated depth.

| Experiments | | | Lock3DFace | | | IIIT-D RGB-D database | | | Our dataset | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Modal | G./P. | Estimated depth | Color | Depth | Fusion | Color | Depth | Fusion | Color | Depth | Fusion |
| $M_1$ | $G_1/P_1$ | - | 94.51% | 79.85% | 95.57% | 95.88% | 78.89% | 96.50% | 93.33% | 94.26% | 98.00% |
| $M_1$ | $G_1/P_2$ | FCN | 94.51% | 1.78% | 94.53% | 95.88% | 4.53% | 95.88% | 93.33% | 11.01% | **93.77%** |
| | | Proposed | | 2.61% | **94.58%** | | 3.81% | 95.88% | | 5.18% | 93.41% |
| | $G_2/P_2$ | FCN | | 38.19% | 94.52% | | 58.19% | 95.98% | | 15.57% | 93.44% |
| | | Proposed | | 45.26% | 94.67% | | 67.04% | **96.19%** | | 17.30% | 93.43% |
| $M_2$ | $G_1/P_2$ | FCN | 94.51% | 1.37% | 94.53% | 95.88% | 4.02% | 95.88% | 93.33% | 1.52% | 93.34% |
| | | Proposed | | 0.93% | 94.53% | | 3.50% | 95.88% | | 0.52% | 93.33% |
| | $G_2/P_2$ | FCN | | 50.50% | 94.28% | | 69.62% | 95.78% | | 52.69% | 94.79% |
| | | Proposed | | 80.07% | **94.90%** | | 82.08% | **96.50%** | | 80.82% | **96.13%** |
| $M_3$ | $G_1/P_2$ | FCN | 94.51% | 21.06% | **94.88%** | 95.88% | 19.77% | 95.98% | 93.33% | 48.87% | 95.84% |
| | | Proposed | | 65.34% | 94.74% | | 25.85% | 96.09% | | 67.82% | **96.49%** |
| | $G_2/P_2$ | FCN | | 48.73% | 94.10% | | 70.44% | 96.09% | | 52.12% | 95.00% |
| | | Proposed | | 65.34% | 94.74% | | 82.60% | **96.40%** | | 81.43% | 96.43% |

Table 5. Face recognition accuracies (rank-1) on the Lock3DFace database.

| Test subset | Description | Only RGB | Depth est. by FCN [24] | | Depth est. by our approach | | Ground-Truth Depth | |
|---|---|---|---|---|---|---|---|---|
| | | | Depth | Fusion | Depth | Fusion | Depth | Fusion |
| Probe_S1 | {NU} x S1 | 100% | 98.51% | 100% | 100% | 100% | 99.55% | 100% |
| Probe_S2 | {FE} x S1 | 100% | 84.55% | 100% | 98.47% | 100% | 98.03% | 100% |
| Probe_S3 | {PS} x S1 | 95.63% | 17.74% | 95.38% | 70.53% | 95.87% | 65.26% | 95.92% |
| Probe_S4 | {OC} x S1 | 97.13% | 39.74% | 97.18% | 78.28% | 97.39% | 81.62% | 99.11% |
| Probe_S5 | {NU, FE, PS, OC} x S2 | 81.56% | 11.66% | 82.44% | 54.27% | 84.53% | 55.79% | 85.13% |
| Total | {S1, S2} | 94.51% | 50.50% | 94.28% | 80.07% | **94.90%** | 79.85% | 95.57% |

is able to improve 2D face recognition performance is that by converting the RGB modality into a new space (i.e., the depth space), the network is able to extract subject discriminative features from a different aspect. These features may not be fully explored in the original RGB image space.

## 5. Conclusions

We propose an end-to-end learning method for estimating the face depth information from a 2D face image, and use the estimated depth information to improve 2D face recognition accuracy. The proposed approach aims retain more subject discriminative information in the estimated depth instead of only minimizing the errors between the estimated depth and the ground-truth depth. Experimental results shown that while the proposed methods achieves similar depth estimation MAE to the state-of-the-art method, its performance of face recognition is much better. In our future work, we will study the effectiveness of the proposed approach for improving 2D face recognition performance using additional modalities such as near-infrared. In addition, face liveness detection [21, 22] and attribute learning [7, 8] based on multi-modality information will be interesting research directions.

## Acknowledgment

This research was supported in part by the Natural Science Foundation of China (grants 61672496, 61390511, and 61732004), National Basic Research Program of China (grant 2015CB351802), External Cooperation Program of Chinese Academy of Sciences (grant GJHZ1843), and the Strategic Priority Research Program of CAS (grant XDB02070004).

## References

[1] Y. Cao, Z. Wu, and C. Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. In *Proc. TCSVT*, number 99, pages 1–1, 2017.

[2] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. *arXiv:1604.03901*, 2017.

[3] A. Chowdhury, S. Ghosh, R. Singh, and M. Vatsa. RGB-D face recognition via learning-based reconstruction. In *Proc. BTAS*, pages 1–7, Sept. 2016.

[4] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv:1406.2283*, 2014.

[5] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh. On RGB-D face recognition using kinect. In *Proc. BTAS*, pages 1–6, Sept. 2013.

Table 6. Face recognition accuracies (rank-1) on IIIT-D and our RGB-D database.

| Databases | Methods | Rank-1 face identification accuracy | | | | |
|---|---|---|---|---|---|---|
| | | Only RGB | Depth | Fusion (score-level) | Fusion (feature-level) | Fusion (4-channel) |
| IIIT-D | FCN | 95.88% | 69.62% | 95.78% | 88.36% | 95.67% |
| | Proposed | | 82.08% | **96.50%** | 96.19% | 95.16% |
| | Ground-truth | | 78.89% | 96.50% | 94.34% | 94.85% |
| Lock3DFace | FCN | 94.51% | 50.50% | 94.28% | 75.65% | 94.25% |
| | Proposed | | 80.07% | **94.90%** | 93.57% | 94.20% |
| | Ground-truth | | 79.85% | 95.57% | 94.31% | 93.74% |
| Our dataset | FCN | 93.33% | 52.69% | 94.79% | 70.58% | 90.90% |
| | Proposed | | 80.82% | **96.13%** | 93.18% | 90.16% |
| | Ground-truth | | 94.26% | 98.00% | 97.47% | 90.52% |

[6] G. Goswami, M. Vatsa, and R. Singh. RGB-D face recognition with texture and attribute features. *IEEE Trans. Inf. Forensics Security*, 9(10):1629–1640, Oct. 2014.

[7] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–13, Aug. 2017.

[8] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(6):1148–1161, Jun. 2015.

[9] H. Han, S. Shan, X. Chen, and W. Gao. A lighting normalization approach exploiting face symmetry. *Journal of Computer Research and Development*, 50(4):767–775, 2013.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, June 2016.

[11] Y. Hua and H. Tian. Depth estimation with convolutional conditional random field network. *Neurocomputing*, 214(Supplement C):546 – 554, 2016.

[12] X. Huang, L. Fan, J. Zhang, Q. Wu, and C. Yuan. Real time complete dense depth reconstruction for a monocular camera. In *Proc. CVPRW*, pages 674–679, June 2016.

[13] T. Huynh, R. Min, and J.-L. Dugelay. An efficient LBP-Based descriptor for facial depth images applied to gender recognition using RGB-D face data. In *Proc. ACCV Workshops*, pages 133–145, 2012.

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015.

[15] D. Kong, Y. Yang, Y. X. Liu, M. Li, and H. Jia. Effective 3D face depth estimation from a single 2D face image. In *Proc. ISCIT*, pages 221–230, Sep. 2016.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, pages 1097–1105. 2012.

[17] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *Proc. 3DV*, pages 239–248, Oct. 2016.

[18] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In *Proc. CVPR*, pages 1119–1127, June 2015.

[19] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proc. CVPR*, pages 5162–5170, June 2015.

[20] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10):2024–2039, Oct. 2016.

[21] K. Patel, H. Han, and A. K. Jain. Cross-database face antispoofing with robust feature representation. In *Proc. CCBR*, pages 611–619, Sep. 2016.

[22] K. Patel, H. Han, A. K. Jain, and G. Ott. Live face video vs. spoof face video: Use of moire patterns to detect replay video attacks. In *Proc. ICB*, pages 98–105, May 2015.

[23] M. Reiter, R. Donner, G. Langs, and H. Bischof. Estimation of face depth maps from color textures using canonical correlation analysis. In *Computer Vision Winter Workshop*. Feb. 2006.

[24] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, April 2017.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[26] Z. L. Sun and K. M. Lam. Depth estimation of face images based on the constrained ICA model. *IEEE Trans. Inf. Forensics Security*, 6(2):360–370, June 2011.

[27] Z. L. Sun, K. M. Lam, and Q. W. Gao. Depth estimation of face images using the nonlinear least-squares model. *IEEE Trans. Image Process.*, 22(1):17–30, Jan. 2013.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, pages 1–9, June 2015.

[29] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. pages 839–846, 1998.

[30] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(9):1226–1238, Sep. 2002.

[31] J. Zhang, D. Huang, Y. Wang, and J. Sun. Lock3DFace: A large-scale database of low-cost kinect 3D faces. In *Proc. ICB*, pages 1–8, 2016.

[32] Y. Zheng and Z. Wang. Robust depth estimation for efficient 3D face reconstruction. In *Proc. ICIP*, pages 1516–1519, Oct. 2008.