Accurate and robust face recognition from RGB-D images with a deep learning approach

Yuancheng Lee
http://cv.cs.nthu.edu.tw/php/people/profile.php?uid=150
Jiancong Chen
http://cv.cs.nthu.edu.tw/php/people/profile.php?uid=153
Ching-Wei Tseng
http://cv.cs.nthu.edu.tw/php/people/profile.php?uid=156
Shang-Hong Lai
http://www.cs.nthu.edu.tw/~lai/

Computer Vision Lab, Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

Abstract

Face recognition from RGB-D images utilizes 2 complementary types of image data, i.e. colour and depth images, to achieve more accurate recognition. In this paper, we propose a face recognition system based on deep learning, which can be used to verify and identify a subject from the colour and depth face images captured with a consumer-level RGB-D camera. To recognize faces with colour and depth information, our system contains 3 parts: depth image recovery, deep learning for feature extraction, and joint classification. To alleviate the problem of the limited size of available RGB-D data for deep learning, our deep network is firstly trained with colour face dataset, and later fine-tuned on depth face images for transfer learning. Our experiments on some public and our own RGB-D face datasets show that the proposed face recognition system provides very accurate face recognition results and it is robust against variations in head rotation and environmental illumination.

1 Introduction

Face recognition from colour face images has been developed for decades, but the recognition accuracy is sensitive to head pose and environmental illumination. In addition, colour-based recognition may be easily deceived by pre-captured video or image. Depth face image, as a complementary type of data to colour face image, contains some superiority at several situations. First, in harsh illumination, depth camera can still provide depth maps by using an active infrared light source. In addition, it can overcome the problem of deceiving. This urges us to consider combining both colour and depth information to develop an accurate and robust face recognition system. Thanks to the recent advances in deep networks, machines can nowadays outperform human on certain colour face recognition problems [2,3,4,5,6]. The success of deep learning is largely due to huge training dataset and GPU computing. How to establish an accurate RGB-D face recognition system based on deep learning remains a challenge.

Pages 123.1-123.14
DOI: https://dx.doi.org/10.5244/C.30.123

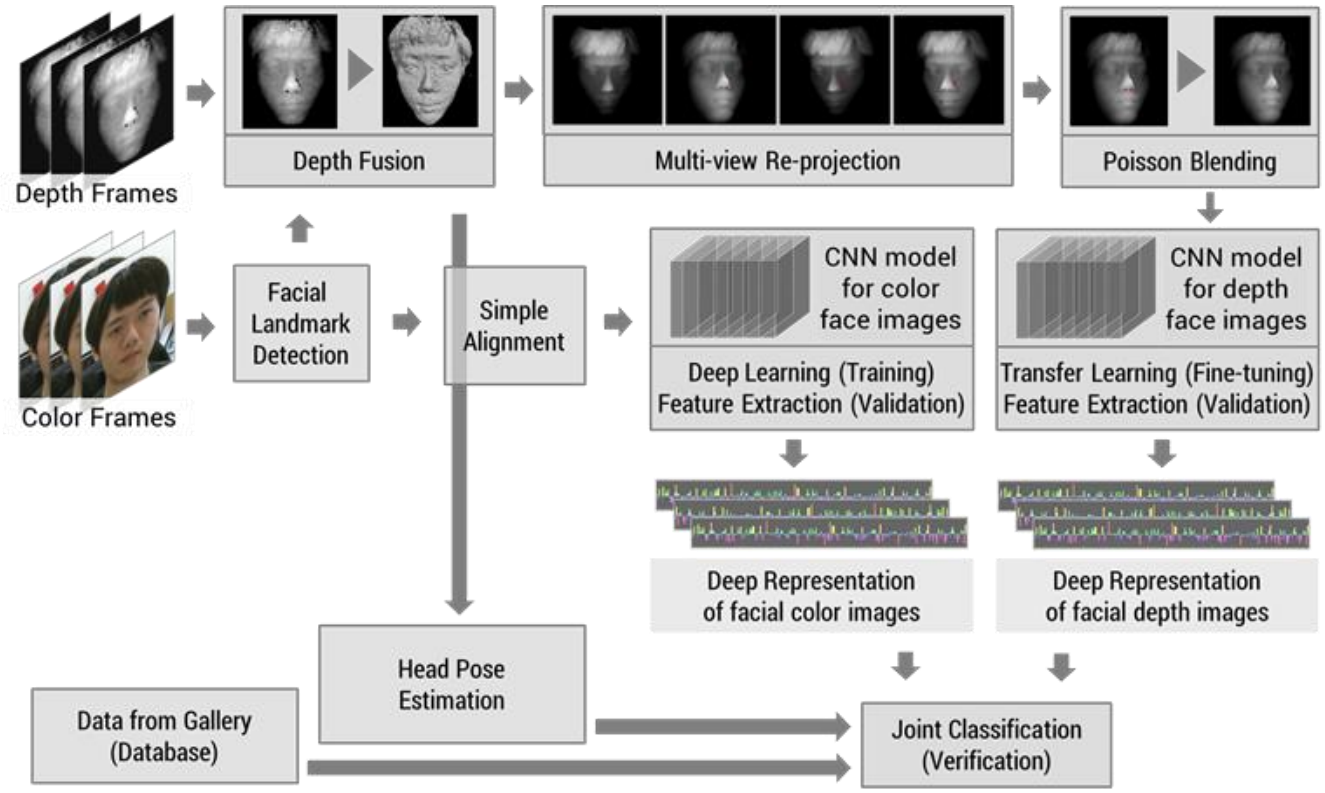


Figure 1: A flowchart of our proposed system. Red region on depth maps means where depth information is lost.

A colour face image represents photometrical sensing of a face under particular illumination condition, which is sensitive to not only head pose and facial expression, but also surface texture and light sources. On the contrary, a depth face image directly acquires 3-D position, so geometry information like shape and surface normal can be easily computed. To conquer the aforementioned weakness of colour-based face recognition system, we aim to exploit complementary and discriminative information from depth face images to achieve more robust face recognition.

In this paper, we propose an efficient pipeline to include 3-D face shape into the deep network for learning effective colour and depth feature transformation. Moreover, in order to design a robust classifier, we further investigate influence on colour-based and depth-based recognition performance of each factor, such as illumination and head pose. After cautious validation, the proposed system has the following contributions:

Depth data enhancement, recovery, and augmentation: A carefully designed pipeline to improve quality of facial depth images from a consumer-level depth camera, including super-resolution, multi-view re-projection, and Poisson Blending [16].

Deep transfer learning: Our model was firstly trained on colour (RGB and grayscale) face images, and later fine-tuned on depth images which helps machine efficiently transfer the knowledge of colour images to depth images.

Joint classifier: To optimally utilize colour and depth information in different situations, an SVM [44,45] with probability estimation takes not only deep representation similarities, but also head pose and database similarity standard deviation (which can be viewed as an index of image quality) into consideration, to jointly estimate a confidence score to make an accurate decision.

2 Related Work

Conventional RGB-D face recognition: We concentrate on the face recognition techniques developed for RGB-D data captured by a consumer-level depth camera, like Microsoft Kinect. Previous works for RGB-D face recognition can be roughly sorted as image-based methods and geometry-based methods. Image-based methods [27,28,29,30] transplant the pipeline used in colour face recognition to depth map for depth face recognition. In [27] and [31], they employed image-level fusion, for their dataset provides precise pixel-level correspondence between colour and depth images. In [28] they proposed to process colour and depth information separately to extract colour and depth LBP features, respectively, while [29] extracts eigenface features from the face images for bi-model classification.

The core concept of geometry-based methods [31,32,33,34,36] is 3-D face model reconstruction for better shape restoration, by using depth fusion or morphology. A depth map can be reprojected into 3D space to generate a point cloud, and further meshed using per-pixel meshing. Borrowing the idea of multi-frame super-resolution, depth fusion is to merge depth information from neighbouring frames to improve the depth quality. Previous methods [31,32,33,34,35,36] implemented depth fusion by point set registration. For real-world application, [31,32] used EM-ICP [32] and [33,34] implemented ICP on GPU for acceleration. For rank-1 recognition, [31,32] randomly select several canonical faces from gallery to simplify the similarity measure computation, while [34] applies SIFT-based descriptor for robust description.

CNN-based Face Recognition: Convolutional Neural Network (CNN) [7] is a well-known end-to-end learning architecture, which is often used to encapsulate and encrypt image data into semantic and abstract representation. For face recognition, two critical problems are considered, face verification and face identification. Face verification is often regarded as a similarity estimation problem. Moreover, a verification algorithm can be easily adapted to solve the identification problem.

In the past two years, several insights about face recognition with deep learning are revealed. Small-size convolution filters and very deep network [21,22] help extract subtle but distinguishable facial features, reduce the total number of parameters, and maximize the nonlinearity of deep architecture. Joint identification-verification supervisory signal [2] spurs the network to learn both aspect of skills. With the assumption that face varieties can usually fit well into low-dimensional manifolds, low-dimensional representation helps lower the number of feature maps, so the scale of network can be reduced. Consequently, recent deep networks evolve with evident structural inheritance.

DeepFace [8], trained on a well labelled face dataset with 4.4 million images, reaching human-level accuracy for the first time, which frontalizes faces in 3D. Softmax was applied for prediction and back-propagation started from cross-entropy loss for each sample. DeepID [1], trained on a much smaller training dataset of 0.2 million images, achieves recognition accuracy as high as DeepFace. The success of DeepID can be attributed to 3 major procedural modifications; namely, data augmentation, ensemble, and statistical classifier. For a deep network, researchers usually take output vector of the last hidden layer as a semantic feature transformation, and further use it as input to a state-of-art classifier, like support vector machine [8,26] or Joint Bayesian classifier [25], to solve the classification problems. Later DeepID series [2,3,5] further refine their work with joint identification-verification supervisory signal, auxiliary supervisory network, very deep architecture, and extended dataset.

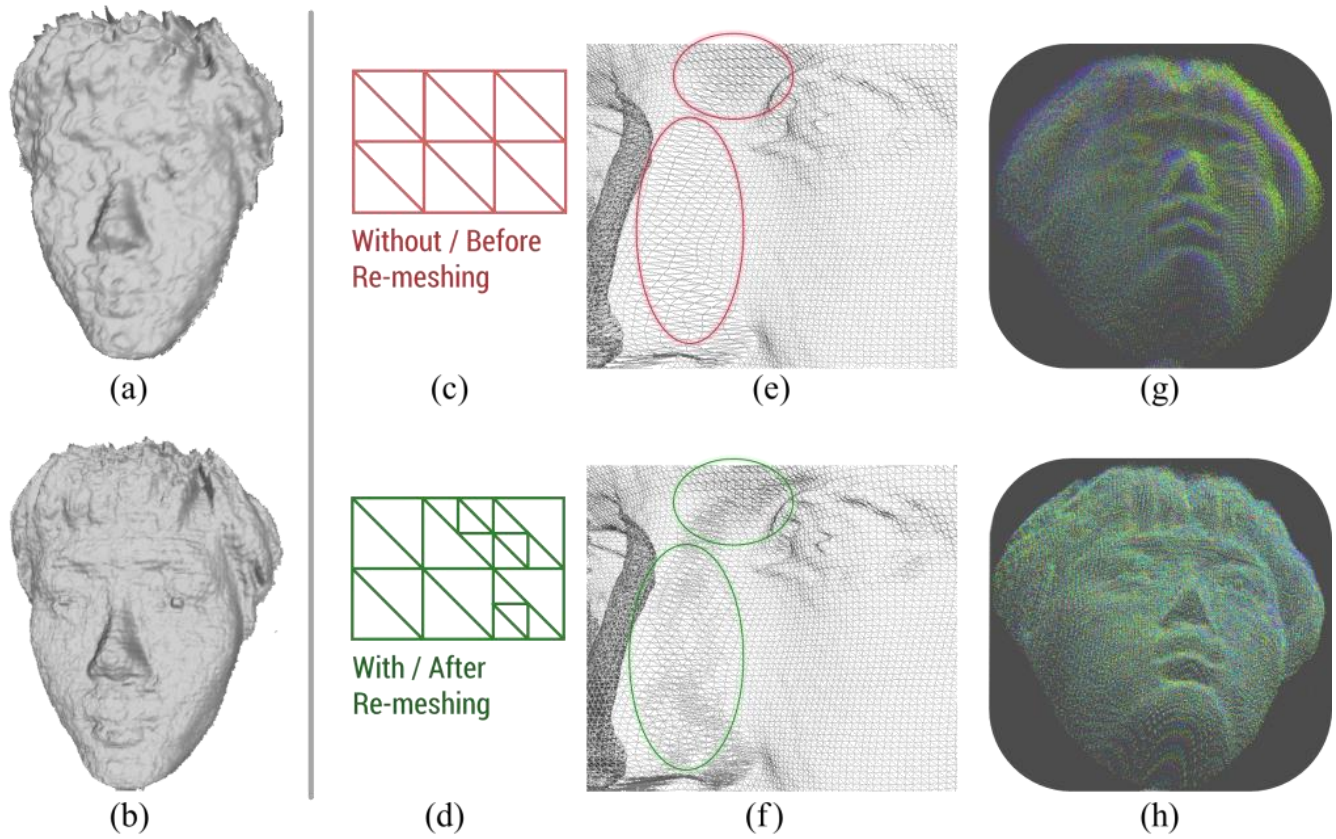


Figure 2: Reconstructed face model from (a) 1 depth frame, and (b) 10 depth frames. Visualization of (c) pixel grid meshing, and (d) re-meshing. Face mesh model (e) before re-meshing, and (f) after re-meshing. Face point cloud registration (g) without, and (h) with re-meshing. Note that point cloud from each frame is rendered in respective colour.

3 Proposed Method

3.1 Depth Face Image Recovery and Enhancement

In this subsection, our pipeline for facial depth data recovery and enhancement is explained step by step. For face positioning, IntraFace [20] was applied to detect facial landmarks on colour image first. Then, with pre-calibrated camera parameters, landmarks can be transformed onto corresponding depth pixels. Our recovery and enhancement pipeline consists of image-based and 3-D geometry-based techniques, so the depth pixels are projected into 3-D space during the pipeline, and rendered back onto images again after a series of processing.

A consumer-level depth camera, such as Kinect, can acquire depth images at high frame rate, but at the cost of poor image quality. Borrowing the concept of multi-frame super-resolution, our depth fusion generates high-quality face shape model (**Fig.** 2(b)) from a sequence of depth frames. Since it is per-pixel depth that captured by a depth camera, directly applying ICP for point set registration (**Fig.** 2(g)) may not be appropriate, for the vertex density is not uniformly distributed all over the point cloud.

For noise reduction, we apply wiener filter [39] to every depth frame with kernel size 3×3. To construct an initial face model from each depth frame, we perform pixel-grid meshing (**Fig.** 2(c)).

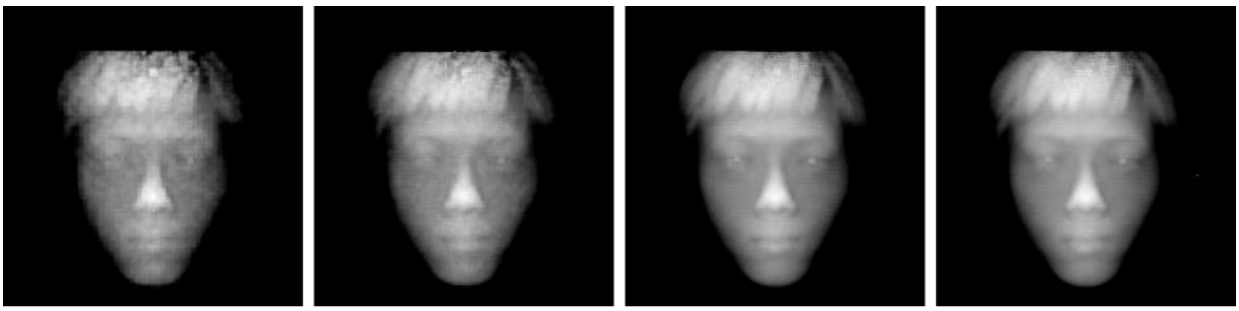


Figure 3: Face Image Recovery and Enhancement using
(a) 1 frame, (b) 5 frames, (c) 10 frames, (d) 30 frames

Re-meshing: Because the density of reprojected point cloud is not uniformly distributed (i.e., where the greater depth gradient, the lower density), point cloud registration may disregard the error caused by those low-point-density regions, like nose. Re-meshing is to make point density consistent all over the face surface.

After projecting a depth image into 3-D space and meshing with pixel grid, we obtain a facial surface model. Let τ_p be 2 times the minimum perimeter of a triangle in the mesh model. For each triangle mesh $\mathbf{M}_i \equiv \Delta(\mathbf{v}_{i1}, \mathbf{v}_{i2}, \mathbf{v}_{i3})$, where $\mathbf{v}_{ij} \in \mathrm{R}^3$, if its perimeter $P(\mathbf{M}_i)$ is longer than τ_p, we divide it into 4 small triangles by adding 3 vertices to middle points of its 3 edges. This process is applied recursively, until there is no triangle with perimeter larger than τ_p. We don't update τ_p during the process (**Fig.** 2(c)(d)).

Coarse-to-Fine Point Cloud Registration: As mentioned previously, ICP is time-consuming and need to be improved for real-world applications. We propose a 2-step coarse to fine registration stratregy, in order to merge a high-quality facial point cloud **Fig.** 2(h) from T low-quality ones. Firstly, for each point set $\mathbf{X}_i \in \mathrm{R}^{\mathrm{N}_t \times 3}$ projected from the t-th depth frame ($t \in \{1, 2, \ldots \mathrm{T}\}$), we estimate a similarity transformation A_t from its 49 facial landmarks $\mathbf{L}_t \in \mathrm{R}^{49\times 3}$ to $\mathbf{L}_{\lceil \mathrm{T}/2 \rceil}$ by

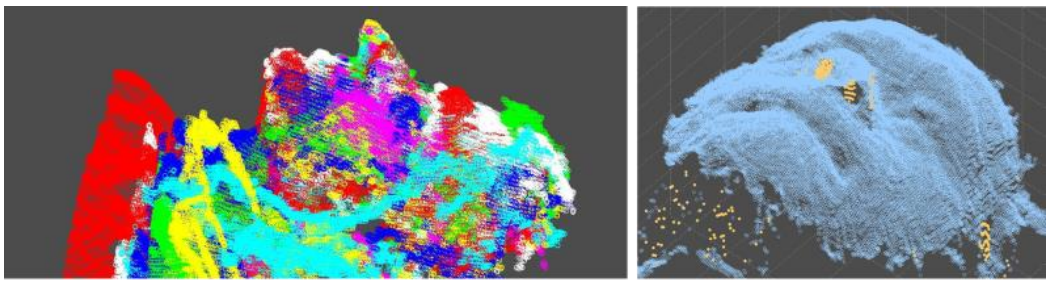
$$argmin_{A_t} \left\| A_{(t,\lceil T/2 \rceil)}(\mathbf{L}_t) - \mathbf{L}_{\lceil T/2 \rceil} \right\|_F^2 \qquad (1)$$

where F is Frobenius norm, and $A_{(t,\lceil T/2 \rceil)}$ is the similarity transformation function from $\mathbf{L}_t$ to $\mathbf{L}_{\lceil T/2 \rceil}$, with parameters estimated by SVD trick.

Secondly, we applied Fast-ICP [41] with Geometrically stable sampling [38] for rigid transformation [40] from $A_t(\mathbf{X}_t)$ to $\mathbf{X}_{\lceil \mathrm{T}/2 \rceil}$ is used to refine A_t to A_t^* . Our high-quality point set $\mathbf{X}^*$ contains all the aligned point set $A_t^*(\mathbf{X}_t)$, ($t \in \{1, 2, \ldots \mathrm{T}\}$). With initial alignment using facial landmark points, refinement (ICP) converges in several (less than 10) iterations. Empirically, T is set to 10 in our implementation. **Fig.** 3 shows the result of enhanced depth map by the subsequent procedure. Recovery and enhancement using 10 frames leads to better quality than using 5 frames (**Fig.** 3(b)(c)). Nonetheless, the quality improvement from 10 frames to 30 frames is negligible (**Fig.** 3(c)(d)).

In addition, face pose can be determined in this phase, since we can estimate similarity transformation from the facial landmarks of referenced frame $\mathbf{L}_{\lceil \mathrm{T}/2 \rceil}$ to an offline computed and frontalized common face $\mathbf{L}_{template}$. Thus, face pose information can be extracted from rotation ingredient of the similarity transformation.

Template, Mean Faces, and Pose Estimation: After depth fusion, there may still be some "holes" (**Fig.** 4(b)), caused by shadow, weak reflection, or occlusion/disocclusion.



(a) (b)

Figure 4: (a) Visualization of template based frontalization of 8 faces (rendered in respective colours). (b) Visualization of hole (rendered in yellow) filling.

For hole filling and face frontalization, common face and mean faces are essential. We define the landmark dissimilarity of 2 facial point sets as the normalized SSD between 2 landmark sets, the aligned one $\mathbf{L}_1$ and the reference one $\mathbf{L}_2$, as follows:

$$Dis(\mathbf{L}_1, \mathbf{L}_2) = \frac{\left\| A_{(\mathbf{L}_1,\mathbf{L}_2)}(\mathbf{L}_1) - \mathbf{L}_2 \right\|_F^2}{\|\mathbf{L}_2 - mean(\mathbf{L}_2)\|_F^2} \tag{2}$$

where $A_{(\mathbf{L}_1,\mathbf{L}_2)}$ is the estimated similarity transformation function from $\mathbf{L}_1$ to $\mathbf{L}_2$, and an affinity matrix is given by:

$$\mathbf{K}_{i,j} = \begin{cases} e^{-\frac{Dis(\mathbf{L}_1,\mathbf{L}_2)}{2\sigma^2}} & , \text{when } Dis(\mathbf{L}_1, \mathbf{L}_2) \le \gamma \\ 0 & , \text{otherwise} \end{cases} \tag{3}$$

Next, the face with maximum average affinity is manually frontalized and taken as facial template. Afterwards, spectral clustering is applied to find 12 common facial "modes" (canonical faces), and mean depth face images of each mode is further computed. By aligning each facial point cloud onto the template using landmark based similarity transformation and decomposing the transformation matrix, we can estimate the capture-time horizontal and vertical rotation of head pose.

3-D Face Model Reconstruction: The mesh structure is destroyed in the point set registration. For super-resolution, we render the frontalized and merged point cloud on a 400×400 canvas, whose resolution is 2 times of the source resolution. In the rendering process, forward bilinear interpolation and Z-buffering is applied, for 3-D to 2-D sampling, and to solve hidden point problem.

There may be some holes on the super-resolution image. With its facial mode estimated before with maximum affinity and mean depth face image computed offline by the same super-resolution strategy, we can fill the holes by Poisson Blending [16], with super-resolution depth image as background, mean depth face image as foreground, and detected hole pixel map as mask. We further reconstruct the 3-D face model by pixel-grid meshing again (**Fig.** 2(b)).

3.2 Learning Deep Representation

Data Augmentation: We implement traditional augmentation tricks, including multiple cropping, mirroring and scale jittering on our aligned colour dataset. For depth data, our augmentation strategy contains scale jittering and multi-view re-projection. Images are essential input for our network, so multi-view re-projection is applied to maximize the superiority of depth data.

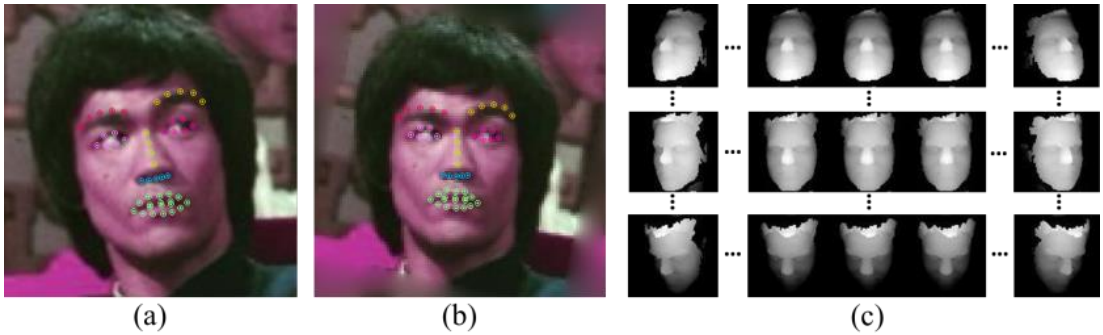


Figure 5: Example of (a), (b) colour face alignment, and (c) multi-view re-projection.

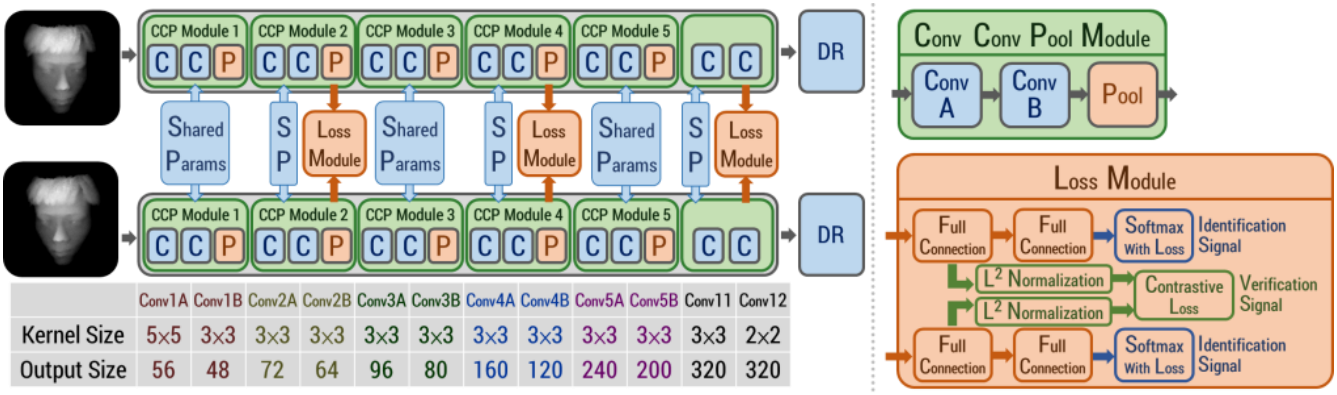


Figure 6: Architecture of our network

For depth data, our augmentation strategy contains scale jittering and multi-view re-projection. Images are essential input for our network, so multi-view re-projection is applied to maximize the superiority of depth data. To synthesize depth maps from different view angles for a single 3-D face model, we rotate the model horizontally and then vertically before rendering. The vertical rotation (pitch) is parameterized by θ_x and the horizontal rotation (yaw) is parameterized by θ_y, where $\theta_y, \theta_x \in \{-25°, -20°, \ldots 0°, \ldots 25°\}$. After each rotation, the model is further rendered onto a depth image $\mathbf{I}_{depth}^{(\theta_y,\theta_x)}$ again with scale jittering. For each 3-D face model, we augment the data by 121(view angles) × 3(scales), that is, 363 images per face.

Deep network: Inspired by the networks proposed by [4] and [5], our deep networks for learning colour and depth deep representation are built with a 12-layer very deep architecture, joint identification and verification supervisory signal, and auxiliary supervisory networks (**Fig.** 6(a))).

The first 5 modules of our network are similar convolution-convolution-pooling (C-C-P) combinations, with max pooling which subsamples feature maps by half. Padding is applied to all the convolutional layers of the 5 C-C-P modules to maintain the size of feature maps. PReLU [48] activation neuron is applied to all the convolutional layers of 5 modules, and random dropout [18] of 0.5 dropout rate is applied to the 11th layer for regularization. The overall parameter number (exclude loss modules) is 2.2 million.

Our network is trained with Caffe [43], an open-source deep learning framework.

Transfer Learning: We first train our network on CASIA-WebFace [4] dataset for colour (RGB and greyscale) face images, which contains 0.5 million labelled face images from 10 thousand subjects. The model for greyscale images is further fine-tuned on the merged and augmented depth dataset of IIR3D [10,11], GavabDB [9], and Texas 3D Face Recognition Database [35], to obtain another deep model for depth face recognition.

3.3 Confidence Estimation

Colour Similarity Estimation: Deep representation (DR) is the output of a high-level fully-connected layer, which can be viewed as a transformed feature. The colour DR of frame t ($t \in \{1, 2, \ldots T\}$) is denoted by $DR\left(\mathbf{I}_{colour}^{(t)}\right)$. The group-wise colour face similarity between 2 group of frames, $\boldsymbol{I}_{color}^{i} \in G_{color}^{1}$ and $\boldsymbol{I}_{color}^{j} \in G_{color}^{2}$, is defined as

$$\sum_{\boldsymbol{I}_{color}^{i} \in G_{color}^{1}} \sum_{\boldsymbol{I}_{color}^{j} \in G_{color}^{2}} \frac{Sim\left(DR(\boldsymbol{I}_{color}^{1}), DR(\boldsymbol{I}_{color}^{2})\right)}{Num(G_{color}^{1}) \times Num(G_{color}^{2})} \tag{4}$$

, where the pair-wise image similarity given by

$$Sim\left(DR(\boldsymbol{I}_{color}^{1}), DR(\boldsymbol{I}_{color}^{2})\right) = Sigmf_{a,c}\left(DR(\boldsymbol{I}_{color}^{1}) \cdot DR(\boldsymbol{I}_{color}^{2})\right) \tag{5}$$

, the $Sigmf_{a,c}$ in which is a sigmoidal membership function parameterized by scale parameter a and mean parameter c, which are learned by coarse-to-fine grid search.

Depth Similarity Estimation: $DR\left(\mathbf{I}_{depth}^{(\theta_x, \theta_y)}\right)$ denotes DR of a re-projected depth image of a certain view angle parameterized by (θ_x, θ_y). Similarity of 2 groups of multi-view depth images is defined as the weighted average of pairwise similarities as follows:

$$\frac{\sum_{(\theta_x^{(1)}, \theta_y^{(1)}, \theta_x^{(2)}, \theta_y^{(2)})} W_{(\theta_x^{(1)}, \theta_y^{(1)}, \theta_x^{(2)}, \theta_y^{(2)})} Sim\left(DR\left(I_{(\theta_x^{(1)}, \theta_y^{(1)})}\right), DR\left(I_{(\theta_x^{(2)}, \theta_y^{(2)})}\right)\right)}{\sum_{(\theta_x^{(1)}, \theta_y^{(1)}, \theta_x^{(2)}, \theta_y^{(2)})} W_{(\theta_x^{(1)}, \theta_y^{(1)}, \theta_x^{(2)}, \theta_y^{(2)})}} \tag{6}$$

where $W_{\left(\theta_x^{(1)}, \theta_y^{(1)}, \theta_x^{(2)}, \theta_y^{(2)}\right)}$ is a pairwise weighting term defined as

$$e^{-\frac{(\sin\theta_x^{(1)} - \sin\theta_x^{(2)})^2 + (\sin\theta_y^{(1)} - \sin\theta_y^{(2)})^2}{2\sigma^2}}, \tag{7}$$

and σ is a scale parameter to be learned.

Joint Confidence Estimation: Our expectation is to achieve robust face recognition. The basic idea is that for different illumination conditions and head pose variations, our classifier can dynamically adjust the weight between colour and depth similarity.

Database similarity standard deviation is proved to be highly correlated to reliability of similarity. For colour images, lower image quality leads to lower database similarity standard deviation. Such a relation can also be found in depth image recognition scenario. Database similarity standard deviation between a query group G_{query} (outside the database) and the database **D** is defined as follows:

$$DBSimStd\left(G_{query}, \mathbf{D}\right) = \sqrt{\frac{1}{\mathrm{N}} \sum_{G_{DB}^{i} \in \mathbf{D}} \left(Sim_{group}\left(G_{query}, G_{DB}^{i}\right) - \mu\right)^2} \tag{9}$$

, where μ is the mean database similarity:

$$\frac{1}{\mathrm{N}} \sum_{G_{DB}^{i} \in \mathbf{D}} Sim_{group}\left(G_{query}, G_{DB}^{i}\right) \tag{10}$$

A support vector classifier [44] with Gaussian kernel and LIBSVM [42] approach for probability estimation is trained with the following pairwise information as features: (i) Group-wise colour similarity, (ii) Group-wise depth similarity, (iii) Average database colour similarity standard deviation of 2 images, (iv) Average database depth similarity standard deviation, (v) Estimated capture-time head pose difference.

4 Experimental Results

Our system is evaluated on 2 public datasets, Eurocom Kinect Face Dataset (EKFD) [12] and SuperFaces [13] Dataset. EKFD is a dataset contains 52 subjects and each contains 2 sessions. Each session of EKFD contains 9 facial conditions, 6 non-occluded of which in are considered: neutral, left and right profile, light on, smile, and open mouth. SuperFaces is a dataset with 20 subjects, each has one session. Every subject is asked to turn his/her head left and right moderately for several seconds. Exactly speaking, 98 to 563 frames are recorded from person to person. For each subject, 9 clips, each of 10 frames are sampled. When sampling, if conditions permit, the median frame of each clip would be critical head pose, that is, leftmost, right most, and front-most, and moreover, 3 clips for left, 3 for right, and 3 for frontal. Note that depth fusion is only applied on SuperFaces.

4.1 Experiment Results on Eurocom Kinect Face Dataset (EKFD)

For verification evaluation, we apply 4-fold cross validation (divided by subject), in which the training fold is for learning joint confidence estimator (probability kernel SVM) and other minor parameters (a, c in **Eq. 4** and σ in **Eq. 7**), and the testing fold is for evaluation. Confidence threshold is decided by equal error rate (EER.)

For identification evaluation, we apply rank-1 recognition and 2×52-fold leave-one-subject out cross validation. We take one of the sessions as gallery at each time, and the other session to train a confidence estimator (SVM with probability estimation). Note that the testing subject is also excluded during training.

Methods, Test settings, Included data, Recognition scenario	Verification Accuracy	Identification Accuracy			
		Rank-1	Rank-2	Rank-3	Rank-4
Ours, Non-occluded, RGB only, 1 vs. 1	99.1%	97.0%	98.0%	99.7%	100%
Ours, Non-occluded, Depth only, 1 vs. 1	90.0%	78.8%	85.9%	87.3%	89.9%
Ours, Non-occluded, RGB and Depth, 1 vs. 1	99.2%	97.6%	99.2%	99.7%	100%
Ours, Non-occluded, RGB only, 6 vs. 6	99.7%	99.0%	100%	100%	100%
Ours, Non-occluded, Depth only, 6 vs. 6	94.2%	80.8%	91.4%	96.2%	97.1%
Ours, Non-occluded, RGB and Depth, 6 vs. 6	99.7%	99.0%	100%	100%	100%
S^2CDL [51], Frontal, RGB and Depth for training, RGB only for testing, 1 vs. 1	---	82.7%	---	---	---
S^2CDL [51], Frontal, RGB and Depth for training, Depth only for testing, 1 vs. 1	---	63.2%	---	---	---
3DLBP and HAOG fusion [52], Frontal & Non-occluded, Depth only, 1 vs. 1	---	97.1%	99.0%	99.0%	99.0%
Efficient SIFT in RGB-D [53], Frontal & Non-occluded, RGB and Depth, 1 vs. 1	---	83.7%	96.6%	96.2%	98.6%

Table 1: Verification and Identification Accuracy on EKFD datasets

4.2 Experiment Results on SuperFaces Dataset

For verification evaluation, we apply 4-fold cross validation (divided by subject). Each pair contains 2 different clips.

For identification evaluation, we apply rank-1 recognition and 3×20-fold leave-one-subject out cross validation. We divide 9 clips of each subject into 3 groups, and each group contains 1 left, 1 frontal, and 1 right clip. 2 groups of each subject are taken as gallery, and the remaining groups are set as probe.

	Verification Accuracy	Identification Accuracy			
		Rank-1	Rank-2	Rank-3	Rank-4
Ours, RGB only, 1 vs. 1	97.5%	93.0%	95.3%	96.7%	98.3%
Ours, RGB only, 10 vs. 10	99.0%	95.0%	97.8%	99.4%	100%
Ours, Depth only (Depth Fusion)	96.9%	90.6%	92.8%	95.6%	97.2%
Ours, RGB and Depth, 10 vs. 10	99.3%	95.6%	97.8%	100%	100%

Table 2: Verification and Identification Accuracy on SuperFaces datasets

By considering both colour and depth information, we can raise the recognition performance, especially on SuperFaces, since the colour image quality in EKFD is very good, and EKFD captures one depth frame each time, so we can't apply depth fusion.

4.3 Experiment Results on Our Dataset

We have established our dataset with Kinect 2, which contains 20 subjects and each has 2 sessions. This dataset contains 5 head poses (Mid, U, D, L, R, that is, looking at camera, elevation of ±15°, and left and right rotation by 15°). 2×20-fold leave-one-subject out cross validation is applied. To realize the impact of harsh illumination, we add "Color--" testing conditions, which means the "bright" clips are excluded.

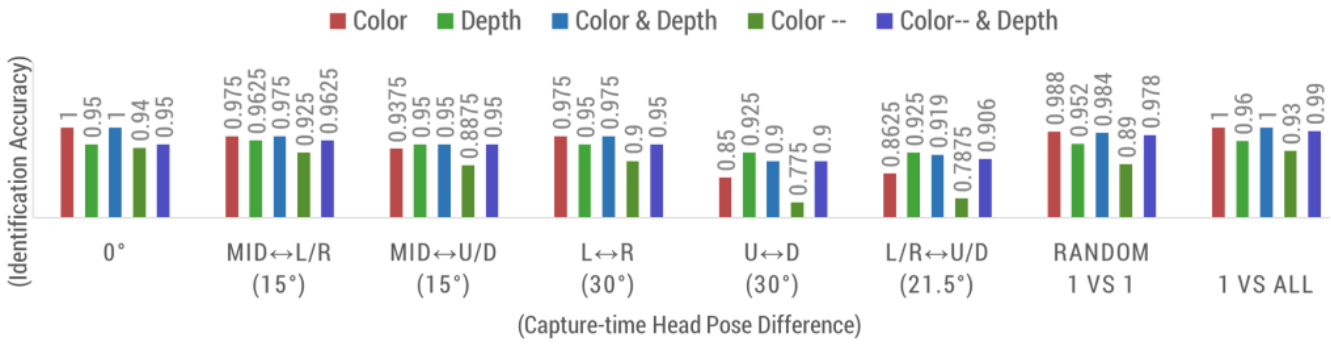


Figure 7: Identification accuracy on our dataset

5 Conclusion

Colour and depth images provide complementary information for face recognition. With deep learning and transfer learning, the classification based on each of them works well independently. Group of frames not only improves depth image quality by depth fusion, but also boosts the performance of face recognition. Our experiments show that higher accuracy can be achieved by using the proposed bi-model confidence estimation, especially under harsh illumination environment or large head pose variation.

References

[1] Y. Sun, X. Wang, and X. Tang, "Deep Learning Face Representation from Predicting 10,000 Classes" *Computer Vision and Pattern Recognition*, 2014.

[2] Y. Sun, X. Wang, and X. Tang, "Deep Learning Face Representation by Joint Identification-Verification" *Conference on Neural Information Processing Systems (NIPS)*, 2014

[3] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust" *Computer Vision and Pattern Recognition*, 2015.

[4] D. Yi, Z. Lei, S. Liao and S. Z. Li, "Learning Face Representation from Scratch" *Computer Vision and Pattern Recognition*, 2015.

[5] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face Recognition with Very Deep Neural Networks" arXiv preprint arXiv:1502.00873, 2015.

[6] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting Ultimate Accuracy: Face Recognition via Deep Embedding" arXiv preprint arXiv:1506.07310.

[7] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification" *Computer Vision and Pattern Recognition*, 2005.

[8] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification" *Computer Vision and Pattern Recognition*, 2014.

[9] A. B. Moreno, and A. Sánchez, "GavabDB: A 3D Face Database," *COST Workshop on Biometrics, on the Internet*, pp. 75-80, 2004.

[10] A. Mian, "Illumination Invariant Recognition and 3D Reconstruction of Faces using Desktop Optics" Optics Express, vol. 19(8), pp. 7491--7506, 2011.

[11] A. Mian, "Shade Face: Multiple Image based 3D Face Recognition", 3D Digital Imaging and Modeling (3DIM)" *Computer Vision-ICCV*, 2009.

[12] R. Min, N. Kose and J. Dugelay, "KinectFaceDB: A Kinect Database for Face Recognition" *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, vol. 44, no. 11, pp. 1534-1548, November 2014.

[13] S. Berretti, P. Pala, and A. Del Bimbo. "Superfaces: A super-resolution model for 3D faces." *Computer Vision–ECCV 2012. Workshops and Demonstrations.* Springer Berlin Heidelberg, 2012.

[14] M. Hernandez, J. Choi, and G. Medioni, "Laser Scan Quality 3-D Face Modeling Using a Low-Cost Depth Camera" *Signal Processing Conference- EUSIPCO*, 2012.

[15] C. Ciaccio, L. Wen, and G. Guo, "Face Recognition Robust to Head Pose Changes Based on the RGB-D Sensor" *(IEEE) Biometrics theory applications and systems*, 2013.

[16] P. Perez, M. Gangnet, and A. Blake, “Poisson Image Editing” *ACM Siggraph*, 2003.

[17] V. Nair, and G. E. Hinton. “Rectified linear units improve restricted boltzmann machines” *Proceedings of the 27th International Conference on Machine Learning (ICML-10),* pages 807–814, 2010.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting” *Journal of Machine Learning Research 15*, pages 1929-1958, 2014.

[19] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014.

[20] X. Xiong, and F. De la Torre, “Supervised Descent Method and its Application to Face Alignment” *Computer Vision and Pattern Recognition*, 2013.

[21] K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition” arXiv preprint arXiv:1409.1556, 2014.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions” arXiv preprint arXiv:1409.4842, 2014

[23] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition” Technical report, arXiv:1409.1556, 2014.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabi-novich. “Going deeper with convolutions” *Computer Vision and Pattern Recognition*, 2015.

[25] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. “Bayesian face revisited: A joint formulation” Proc. *European Conference on Computer Vision*, 2012.

[26] L. Wolf and N. Levy. “The SVM-minus similarity score for video face recognition” *Computer Vision and Pattern Recognition*, 2013.

[27] P. Xiong, L. Huang, and C. Liu. "Real-time 3D face recognition with the integration of depth and intensity images" *Image Analysis and Recognition. Springer Berlin Heidelberg*, 2011. 222-232.

[28] A. Aissaoui and J. Martinet, “Bi-modal face recognition - How combining 2D and 3D clues can increase the precision” *VISAPP*, 2015

[29] F. Tsalakanidou, D. Tzovaras, and MG. Strintzis. "Use of depth and colour eigenfaces for face recognition" *Pattern Recognition Letters 24.9 (2003)*: 1427-1435.

[30] R. Min, et al. "Real-time 3D face identification from a depth camera" *Pattern Recognition (ICPR), 2012 21st International Conference* on. IEEE, 2012.

[31] GS J. Hsu, YL Liu, HC Peng, and PX Wu "RGB-D-based face reconstruction and recognition" *Information Forensics and Security, IEEE Transactions* on 9.12 (2014): 2110-2118.

[32] J. Choi, A. Sharma, and G. Medioni. "Comparing strategies for 3D face recognition from a 3D sensor" *RO-MAN, 2013 IEEE*. IEEE, 2013.

[33] S. Berretti, P. Pala, and A. Del Bimbo. "Increasing 3D Resolution of Kinect Faces." *Computer Vision-ECCV 2014 Workshops.* Springer International Publishing, 2014.

[34] S. Berretti, P. Pala, and A. Del Bimbo. "Face Recognition by Super-Resolved 3D Models From Consumer Depth Cameras." *Information Forensics and Security, IEEE Transactions on* 9.9 (2014): 1436-1449.

[35] S. Gupta, K. R. Castleman, M. K. Markey and A. C. Bovik, "Texas 3D Face Recognition Database" URL: http://live.ece.utexas.edu/research/texas3dfr/

[36] C. Ciaccio, L. Wen, and G. Guo. "Face recognition robust to head pose changes based on the RGB-D sensor." *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on.* IEEE, 2013.

[37] S. Granger and X. Pennec. "Multi-scale EM-ICP: A fast and robust approach for surface registration." *Computer Vision-ECCV 2002* (2006): 69-73.

[38] N. Gelfand, et al. "Geometrically stable sampling for the ICP algorithm." *3-D Digital Imaging and Modeling, 2003. 3DIM 2003. Proceedings. Fourth International Conference on. IEEE*, 2003.

[39] J. Chen, et al. "New insights into the noise reduction Wiener filter." *Audio, Speech, and Language Processing, IEEE Transactions on* 14.4 (2006): 1218-1234.

[40] T. Jost and H. Hugli. "A multi-resolution ICP with heuristic closest point search for fast and robust 3D registration of range images." *3-D Digital Imaging and Modeling, 2003. 3DIM* 2003. Proceedings. Fourth International Conference on. IEEE, 2003.

[41] S. Rusinkiewicz, and M. Levoy. "Efficient variants of the ICP algorithm." *3-D Digital Imaging and Modeling, 2001. 3DIM* 2001 Proceedings. Third International Conference on. IEEE, 2001.

[42] CC Chang and CJ Lin. "LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology." 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[43] Y. Jia, et al. "Caffe: Convolutional architecture for fast feature embedding." Proceedings of the ACM International Conference on Multimedia. ACM, 2014.

[44] BE Boser, IM Guyon, and VN Vapnik. "A training algorithm for optimal margin classifiers." Proceedings of the fifth annual workshop on Computational learning theory. ACM, 1992.

[45] N. Das, D. Mandal, S. Biswas. "Simultaneous Semi-Coupled Dictionary Learning for Matching RGBD Data." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016.

[46] JB C. Neto and AN Marana. "3DLBP and HAOG fusion for face recognition utilizing Kinect as a 3D scanner." *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. ACM, 2015.

[47] MI Ouloul, et al. "An Efficient Face Recognition Using SIFT Descriptor in RGB-D Images." *International Journal of Electrical and Computer Engineering* 5.6 (2015).

[48] K. He, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.