Mixnet Face Recognition How Combing 2D and 3D Data Can Increase the Precision

Xinyu Zhang\textsuperscript{a}, Yang Zhao\textsuperscript{b}, Hao Zhang\textsuperscript{c}
Department of Electronic Engineering Tsinghua University Beijing, 100084, P.R. China
\textsuperscript{a}xy-zhang17@mails.tsinghua.edu.cn, \textsuperscript{b}zhao-yan18@mails.tsinghua.edu.cn, \textsuperscript{c}haozhang@tsinghua.edu.cn

Abstract. Great breakthrough has been made on Classic 2-D face recognition in normal environment. The facenet [1] proposed by Google in 2015 has 99.65% accuracy on LFW dataset. It can be considered that the performance of traditional 2-D network in normalized environment such as security check and bank is reliable enough. However, the accuracy of 2-D face recognition will significantly reduce when processing angular faces or occluded face. This paper claims that the lack of depth information is the reason why the recognition ability of neural network is inferior to that of human beings under complex conditions. Furthermore, we have proposed a method, which can obtain 100\% accuracy on both Texas and Bosphorus dataset by using both 3-D deepmap and 2-D rgb picture as input.

1. Introduction
Face recognition in unconstrained images is the forefront of algorithm perception revolution. Indeed, the popularization of unconstrained face recognition technology will not only bring broader application prospects for face recognition technology, but also enter more common life situations from formal application scenes. Faceid of Iphone is a typical example. In this process, the robustness of face recognition technology must be improved. The current state of art 2-D face recognition technology performs well in standard data sets or strictly set application scenarios. However, its performance is deceased sharply in the case of an unsatisfactory environment (angular face and occlusion)
We humans can obtain distance information with depth
Through binocular vision. In many areas, such as object detection, computer vision has a stronger ability than human beings, which indicates that the neural network architecture of computer vision itself is mature enough to solve visual problems. We believe that giving the neural network input information as plentiful as that of human beings can make it work better. To imitate human visual process, we combine 2-D rgb pictures and 3-D deepmap as input, indicates that besides traditional 2-D network we have to work out with a bran-new 3-D face recognition network.
Compared with 2-D networks, 3-D face recognition networks have a major difficulty - smaller size of data sets. Data collection is relatively easier for 2-D network since 2-D digital camera has been widely popularized. In addition, vast amounts of users’ photos with labels can be produced by social media such Facebook and Twitter. However, for 3-D network, 3-D sensors are much more expensive and less popular. Additionally, professional assistance is needed when taking 3-D deep images. These
lead to less samples and scans in 3-D datasets, for example, Bosphorus Dataset [2] has only 4000+ scans and Texas Database [3] has just 4000+scans, making it difficult for 3-d researchers to find suitable and enough public dataset. Furthermore, small data sets tend to cause overfitting when training the network. The performance of face recognition network largely depends on the size of data set. Google’s facenet has optimized the network performance with a large dataset of 200M+ scans.

What's more, the accuracy of 3-D structured light camera is inversely proportional to the square of distance, which means that the accuracy of structured light camera is poor at a relatively long distance. Due to technical constraints, the resolution of purchasable structured light camera is much lower than that of ordinary RGB camera. Generally speaking, the quality of 3-D depth maps is much lower than that of ordinary color photographs, making 3-D face recognition much more difficult.

We tested the classic Google facenet and attempted to train our own 3-D network. It is found that the traditional 2-D face recognition network under the standard application environment is better than our 3-D network, while our 3-D network has a higher recognition accuracy in the case of partial occlusion and angular face. In our view, the 3-D to 2-D projection of the face produces serious distortion, which results in the difference between the feature extraction results of 2-D network from angular and front face. By contrast, the 3-D depth map can keep the information of relative distance between the pixel points. As a result, although deepmap of angular face seems different, the 3-D network can still get some invariant distance information. This makes 3-D face recognition almost angle immunity. Meanwhile, because the depth change exists widely in human face, the network can still extract features almost equally on all the face area, which means enough features can still be obtained for classification, even occlusion happens on the key areas like eyes. The argument will be proved in subsequent experiments of this paper.

We hope to propose a face recognition method, which can not only introduce 3-D information to reduce the recognition rate decline caused by side face and occlusion, but also preserve the high accuracy of traditional 2-D face recognition under normal conditions.

Based on the above views, mixnet is proposed. Mixnet is a hybrid classifier based on features extracted from rgb images and deepmap. By conducting experiment we proved that mixnet do have a better performance than both our 3-D network and traditional 2-D face recognition network.

Our main contributions can be summarized as follows: (i) we increased the recognition accuracy on Bosphorus dataset to 100%. (ii)We proposed a method to combine 2-D and 3-D information for face recognition, which was proved useful. (iii)We obtained a 3-D face recognition network with high performance. (iii)We explained why the introduction of 3-D information will improve accuracy using theoretical analysis and experiments.

2. Related work
In 2007, Mian Ajmal [4] first combined 2-D and 3-D information for face recognition. With the introduction of neural network algorithm, the accuracy of face recognition has been greatly improved, making the combining method seem uncompetitive. Although the methods using simple 2-D rgb picture as inputs achieve a high accuracy, they still have to confront the above problems - angular face and occlusion.

Deepface [5] is the representative method of traditional 2-D network response to angular face. For 2D angular face, it identifies and locates the key points and then conducts rotating projections on a general 3D model with same key-points marked and tracked to form a series of key-point 2-D location map with angle information. Through evaluating the similarities among these key-points 2-D location maps (real data gain from the testing image and artificial informative data), they estimate the camera-shoot-angle of the picture, and then rotate the 2D angular face to a 2D positive face. In fact, the key-points of the recognizing face and the 3-D general face do not coincide completely, which may affect the angle estimation. In addition, 2-D face rotation has to consider the occlusion of the face itself, such as the part covered by the nose. Consequently, this results in the loss of information and the decline of image quality.
Deepid [6], the traditional method of 2-D network which divides the picture into 60 regional training networks in advance, can respond to occlusion. Besides, they expect that there will be some areas where the occlusion is removed and meanwhile facial information is retained as much as possible.

In practical applications, we find that 2-D network relies heavily on eye information, meaning that different parts of the face are not equivalent to 2-D network. If occlusion occurs in the eyes, such as sunglasses, it is difficult for 2-D networks to obtain reliable recognition accuracy through the remaining information.

3. Mixnet

3.1. Architecture

Mixnet is a composite network system based on coaxial depth input and RGB input - combines 2-D and 3-D networks. The feature vectors extracted from 2-D and 3-D networks are input into the hybrid classifier simultaneously and together they affect the classifier training process and recognition results.

![Fig. 1 The architecture of mixnet.](image)

3.2. Inception-resnet

In 2-D network, we use Google's facenet. The facenet model is chosen because facenet has the largest amount of training data in the current network. At the same time, facenet does not do special optimization for angular face and occlusion. It entirely counts on training with comprehensive and large quantity of data to solve all the problems. Trained with enhanced occlusion and angular face datasets, the proposed 3-D network has a better robustness under poor using conditions. However, because 3-D features are less obvious than those of 2-D and 3-D face has less datasets, the performance of our 3-D network on front-unoccluded face is worse than mature 2-D Network. The combination of these two networks is conductive to achieving better results.

We use inception-resnet as the backbone of our 3-D network and reprogram the convolution layer and the full connection layer according to the size of the images we input. Because the size of the input image is small (160*160), we reduce the convolution layer to five layers, in which each layer uses 1*1 convolution kernel and 3*3 convolution kernel. In the process of reducing the convolution layer, it will be enough as long as the number of convolution layers and pooling layers ensures that in the final convolution result, every pixel in the result layer obtained by 3*3 convolution kernel is related to any pixel of the original image.
### Tab. 1 The architecture of 3-d network

| LAYER | SIZE-IN     | SIZE-OUT    | KERNEL1 | KERNEL2 |
|-------|-------------|-------------|---------|---------|
| CONV1 | 160*160*1  | 40*40*64   | 7*7*1   |         |
| CONV2 | 40*40*64   | 20*20*192  | 1*1*64  | 3*3*64  |
| CONV3 | 20*20*192  | 10*10*384  | 1*1*192 | 3*3*192 |
| CONV4 | 10*10*384  | 10*10*512  | 1*1*384 | 3*3*384 |
| CONV5 | 10*10*512  | 10*10*512  | 1*1*512 | 3*3*512 |
| POOL  | 10*10*512  | 5*5*512    |         |         |
| FC1   | 5*5*512    | 8192*1     | Maxout p=2 |
| FC2   | 8192*1     | 8192*1     | Maxout p=2 |
| FC3   | 8192*1     | 512*1      |         |

### 3.3. Network optimization
Because the public 3-D photo data set is collected by laser scanning, it has high quality and preserves most of the facial details intact. When we set the weight of 1*1 and 3*3 convolution kernels of each convolution layer, we treat the overall and local features equally, so the number of 1*1 convolution kernels and 3*3 convolution kernels are equal. However, in practical application, we use structured light camera to collect data and the qualities are poor. Besides, with the increase of distance, the image quality will be seriously attenuated. To this end, we modified the ratio of 1*1 and 3*3 convolution kernels in the second and third convolution layers to 1:2, and depend more on the overall features (in most pictures, the area covered by the second and third convolutions is roughly the size of a facial organ).

![Fig. 2 Picture of Texas Dataset and Picture taken by structured light camera](image)

### 4. Experiment

#### 4.1. Enhanced training set

|          | Original Texas | Original Bos | Rotated Texas |
|----------|----------------|--------------|---------------|
| Class    | 118            | 101          | 118           |
| Scans    | 1149           | 4426         | 11490         |
| Texas occlusion | 118 | 101 | 12 |
| Bos occlusion | 4426 | | 1200 |
| Obtained Pic | | | 74015 |

Texas dataset contains scans with no occlusion, little unimportant expressions and all front face. We rotate the scans randomly, and add random occlusion to simulate the images obtained by the face recognition system under poor using conditions. Meanwhile, the data sets are expanded. We do not rotate the BOS data set because it contains angular face photos with complete angle information.
Although the BOS data set itself contains some scans with hand-shaded area and glasses, but the number is not enough. Hence, we artificially add some random occlusion. In addition, we add some data collected by structured light camera (mainly as verification set).

On this basis, we not only add a random bias to the overall depth on all data sets, but also do some random scaling to the relative depth to prevent over-fitting.

In the experiment, we divided the expanded data set into 10% classes as test set and the rest as training set. We use the training set composed of 90% classes to train feature extraction network, and verify the recognition accuracy on the test set. We repeat this process ten times to ensure that each class in the dataset is included once in the test set. When calculating the recognition accuracy, we only count the original pictures of the original data set to ensure the comparability of the recognition rate.

Expanded dataset

![Expanded dataset](image)

**Fig. 3** expanded dataset

### 4.2. Partial occlusion recognition

To test the robustness of 3-D network to occlusion, and compare with 2-D network, we put top-down and top-down occlusion of different size on all Texas dataset to study the decline of classification accuracy with the increase of occlusion area.

![Top-down occlusion remaining area 0.5 0.6 0.6(rgb)](image)

**Fig. 4** top-down occlusion remaining area 0.5 0.6 0.6(rgb)

![Down-top occlusion remaining area 0.5 0.6 0.6(rgb)](image)

**Fig. 5** down-top occlusion remaining area 0.5 0.6 0.6(rgb)
For top-down occlusion, the performance of 3-D network is much better than that of traditional 2-D facenet. Obviously, when the remaining area is less than 0.7, the eye parts in RGB images will be occluded, which corresponds to common occlusion such as glasses and sunglasses in practical applications. With the absence of eye information, the recognition accuracy of facenet decreases significantly, which means the traditional 2-D network relies heavily on eye information. In the 68 key-points of dlib[7], 12 are in eyes area and 10 are in eyebrows area. Compared with the whole face, the information around the eyes contribute a relatively larger weight. According to the down-top occlusion results, we find that although 3-D network is less sensitive to eye information than 2-D network (eyes and eyebrow’s depth characteristics are not as obvious as their rgb characteristics), 3-D network can still extract enough information and get almost the same accuracy as facenet. This indicates that mixnet will have better occlusion robustness than traditional 2-D network.

4.3. Angular face recognition
To recognize side-face data, we test the ability of the network to recognize angular face data by selecting images with angles from the BOS data set.

| Classifier | Training set | Side face | Front face | Side & Front face |
|------------|--------------|-----------|------------|------------------|
| facenet    |              | 0.937     | 0.947      | 0.963            |
| 3-d network|              | 0.997     | 1.000      | 1.000            |
We selected total 7*101 scans all of angular face as test data. Obviously, compared with 2-D network, 3-D network has significantly improved the ability of angular face recognition. In our view, although shooting from different unknown angles, due to the preservation of depth information, we can still restore the relative distance between points, forming an invariable 3-D space point cloud. This information is enough for further recognition. The network composed of convolution layer and full connection layer can imitate the operation of rotation matrix. Using different convolution kernels to represent all possible spinning angles, 3-D networks can extract the same feature map from deepmap taken with different angles. Additionally, we add a large number of positive sample pairs of a person's front face and angular face in the training set, which is conducive to the convergence of the neural network in this direction, making 3-D network angle immunity.

![Fig. 8 Point Cloud in Bos Data Set](image)

For 2-D networks, data with both front and side faces can be added to the training set to improve the classification results. However, the features extracted from the front face and the angular face may be totally different. Due to the redundant dimension of the high-dimensional classifier, the classifier can find a collection classification area including these totally different feature vectors.

In facenet's original paper the authors reported that the validation set accuracy will have the best value when the feature vector dimension is 128, while the performance will decrease if it increases. In our view, the reason is that for the same person's face and angular face in the data set, the network cannot extract exactly same features, leading to the increase of the classification area. The high-dimensional classifier can easily deal with more classification areas in the training set. However, these will cause a degradation of classifier performance with large amounts of testing data.

4.4. Accuracy on public dataset
After training the 3-D network, we merge the feature vectors extracted by facenet and 3-D network, and then input them into the classifier of higher dimensions. Furthermore, we test the classification results on open datasets and compare them with well-known methods in the Field.
Tab. 4 Accuracy on Texas and Bos dataset and comparison

| METHOD              | MODALITY | PUBLIC DATASET |          |          |
|---------------------|----------|----------------|----------|----------|
|                     |          | Texas          | Bosphorus|          |
| VGG-FACE[8]         | 2-d      | 0.997          | 0.964    |          |
| 3-D KEYPOINT LI[9]  | 3-d      | 0.966          |          |          |
| R3DM[10]            | 3-d      |                | 0.981    |          |
| FACENET             | 2-d      | 1.000          | 0.987    |          |
| 3-D THIS PAPER      | 3-d      | 0.997          | 0.993    |          |
| MIXNET              | 2-d+3-d  | 1.000          | 1.000    |          |

Texas contains front unoccluded face only with can represent the normalized using conditions while the Bosphorus dataset is much more complicate with expressions angular face and occlusions.

When classifying the mixed feature vectors combined by 2-D and 3-D vectors, we use a higher-dimensional classifier. In order to prove that the higher recognition rate is because of the better clustering of extracted features, rather than higher-dimensional classifiers.

We randomly find six classes in bos dataset and work out their feature vectors extracted by facenet and mixnet separately.

Fig. 9 Distance between facenet feature vectors

Fig. 10 Distance between mixnet feature vectors
We can see that the distance between face pairs of different classes extracted by mixnet is larger than that of facenet, which means the classifier can distinguish different classes much easier.

It is worth mentioning that 2-D and 3-D networks both have some drawbacks. For example, 2-D networks have poor accuracy on bos dataset because of angular and occlusion. Although 3-D network can almost immune the interference caused by angle, it lacks important color information, which is much more plentiful than deep features. Besides, the 3-D network is trained on a relatively smaller dataset. Both these lead to the recognition accuracy of 3-D network is lower than 2-D network on Texas dataset. By combining the features extracted from 2-D images and 3-D depth maps, we obtain a new feature vector with more information. These new feature easier to be classified. As a result, we improve the accuracy of both Texas and BOS datasets to 100%.

5. Conclusion
In this paper, we have proposed a hybrid network model called mixnet. Imitating human visual input, we simultaneously extract features from 2-D and 3-D deep map for face recognition increasing the recognition accuracy on Bosphorus dataset to 100%. Through conducting separate experiments, we found that the 3-D network is almost immune to rotational deformation and less sensitive to occlusion in key areas, making it more reliable in poor using conditions. We have laid a theoretical foundation for 2-D&3-D hybrid face recognition based on neural network and proved the superiority of hybrid face recognition. Compared with the traditional 2-D face recognition field, the 3-D field receives less attention. However, 2-D face recognition has many difficulties making it hard to be widely used especially in daily-life applications. We believe that with the popularity of 3-D sensors, hybrid face recognition combining 2-D and 3-D information will be more accepted by face recognition researchers and application developers.

References
[1] Schroff, Florian, D. Kalenichenko, and J. Philbin. "FaceNet: A Unified Embedding for Face Recognition and Clustering." 2015.
[2] S. Gupta, M. K. Markey, and A. C. Bovik. Anthropometric 3D face recognition. International journal of computer vision, 90(3): 331–349, 2010.
[3] A. Savran, N. Alyüz, H. Dibeklioğlu, O. C. eliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3D face analysis. In Biometrics and Identity Management, pages 47–56. Springer, 2008.
[4] A. Mian, M. Bennamoun, and R. Owens. An efficient multimodal 2D-3D hybrid approach to automatic face recognition. IEEE TPAMI, 29(11): 1927–1943, 2007.
[5] Taigman, Yaniv, et al. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification." 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE Computer Society, 2014.
[6] Sun, Yi, X. Wang, and X. Tang. "Deep Learning Face Representation from Predicting 10,000 Classes." 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE Computer Society, 2014.
[7] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In BMVC, page 6, 2015.
[8] Kazemi, Vahid, and J. Sullivan. "One Millisecond Face Alignment with an Ensemble of Regression Trees." 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE Computer Society, 2014.
[9] Li, Huibin, et al. "Towards 3D Face Recognition in the Real: A Registration-Free Approach Using Fine-Grained Matching of 3D Keypoint Descriptors." International Journal of Computer Vision 113.2(2015):128-142.
[10] Gilani, Syed Zulqarnain, A. Mian, and P. Eastwood. "Deep, Dense and Accurate 3D Face Correspondence for Generating Population Specific Deformable Models." Pattern Recognition (2017): S0031320317301644