Toward a Robust Shape and Texture Face Descriptor for Efficient Face Recognition in the Wild

Rahma Abed(✉), Sahbi Bahroun, and Ezzeddine Zagrouba

Laboratoire LIMTIC, Institut Supérieur d'Informatique, Université de Tunis El Manar, 2 Rue Abou Rayhane Bayrouni, 2080 Ariana, Tunisie
{rahma.abed,sahbi.bahroun}@isi.utm.tn, ezzeddine.zagrouba@uvt.tn

Abstract. Face recognition in complex environments has attracted the attention of the research community in the last few years due to the huge difficulties that can be found in images captured in such environments. In this context, we propose to extract a robust facial description in order to improve facial recognition rate even in the presence of illumination, pose or facial expression problems. Our method uses texture descriptors, namely Mesh-LBP extracted from 3D Meshs. These extracted descriptors will then be used to train a Convolution Neural Networks (CNN) to classify facial images. Experiments on several datasets has shown that the proposed method gives promising results in terms of face recognition accuracy under pose, face expressions and illumination variation.

Keywords: Face recognition · Mesh-LBP · Convolution neural networks · 3D morphable model

1 Introduction

Face recognition is the most effective technique and one of the most widely used biometrics for identifying and verifying people compared to voice, fingerprints, iris, retina, eye scanner, gait, ear and hand geometry [1]. However, face images suffer from several issues that could affect the achieved results, especially in an unconstrained environment. Such as facial expression, aging, accessories or even occlusion, low resolution, noise, illumination and pose variation [2]. Recently, several deep learning based face recognition methods was proposed [38]. These methods offer promising results in controlled environments. However, these results significantly decrease in real life scenarios.

In order to enhance face recognition results under these issues, two alternatives are offered. Face frontalization or robust face feature extraction. Face frontalization aim to produce a new face image, neutral and frontal, from the original image [3]. While robust face feature extraction extract a discriminative face representation using one or various face feature extractors. Nevertheless, these technique seems to be highly complex as the learning process requires a considerable amount of time and a large dataset for training [39].

N. Tsapatsoulis et al. (Eds.): CAIP 2021, LNCS 13053, pp. 319–328, 2021.
https://doi.org/10.1007/978-3-030-89131-2_29

For this purpose, we propose to use both face feature extraction and deep learning techniques in order to build a robust face recognition system. For this aim, we propose to use shape model and texture descriptor to obtain a robust face feature descriptor against facial expression, pose and illumination. Afterwards, we train a Convolution Neural Network (CNN) model for efficient facial recognition.

2 Related Works

Many face recognition algorithms still face difficulties when it comes to identify faces in large pose face images. These challenges have become a key factor that limit the effectiveness of face recognition in unrestricted environments [40].

Several techniques are used in order to enhance face recognition results. In this work, we focus on multi-modal 2D/3D and deep learning based methods. The multi modal techniques take benefit of the 3D face texture and the 2D face image descriptors to improve the recognition rate by considering the 3D face modeling as an intermediate step for 2D face recognition.

To deal with facial expressions issues, Abbad et al. [25] propose a 3D face recognition system based on feature extraction using geometric and local shape descriptors. Deng et al. [27] employed different features extraction based on local covariance operators. Zhang et.al [26] propose a data-free method for 3D face recognition using generated data from Gaussian Process Morphable Models (GPMM). Recently, Koppen et al. [31] propose a Gaussian mixture 3D morphable face model (GM-3DMM) that models the global population as a mixture of Gaussian subpopulations, each with its own mean, with shared covariance. These models are is constructed using Caucasian, Chinese and African 3D face data.

On the other hand, deep learning techniques train a deep model in order to predict the correct identity of the face image fed as input. FaceNet [36] use a deep convolutional network and maps a face images into a compact Euclidean space where distances correspond to a measure of face similarity. Parkhi et al. [19] fuse a very deep convolution neural network and the triplet embedding for building a robust face recognition system named VGG-faces. Wen et al. [18] propose a center loss function to estimate distance between images. Deng et.al [20] propose the measure of Deep Correlation Feature Learning (DCFL) for measure the correlation loss, which lead to create a large correlation between the deep feature vectors and their corresponding weight vectors in softmax loss. In correlation loss, it applies a weight vector in softmax loss as the prototype of each class.

In this work, we propose a new method based on fusing feature descriptors and 3D model with a neural network. We perform face feature extraction from a detected and aligned 3D face data using mesh-LBP. Indeed, the use of 3D data aim to reduce the impact of pose variation in facial image. In addition, when using mesh-LBP, we obtain a robust descriptor against pose, illumination and facial expression variation, which is not as expensive as generating new face image from a 3D model. Then, the obtained features will be fed into a neural network Face recognition. In our method we use raw images as our representation. We also provide a new CNN architecture through the use of the **locally connected layer**. This network will be trained on a very large labeled dataset.

3 Proposed Method

The proposed method is composed of three steps: Face detection and landmarks location, face feature extraction and CNN training. More details are shown in the following section.

3.1 Face Detection and Landmarks Location

In order to detect and crop facial region from images, we use the Dlib face detector [4]. As well as the detection, Dlib also performs face landmarks localization. This localization is very useful for extracting the most important facial structures from a face image. The Dlib face detector works as follows. First, the face detection and location. Then, the landmarks detection occurs. We highlight that the major facial areas to be labeled are the mouth, right eye and eyebrow, left eye and eyebrow, nose and jaw. The landmarks are provided as 68 point pairs (x, y) that correspond to the labeled facial areas.

3.2 Face Feature Extraction

Our method use 3D data obtained by the use 3D Morphable Face Models (3DMM) [5]. The 3D data could be used as an intermediate step to enhance 2D face recognition performance by modelling the difference in the texture map of the 3D aligned input and reference images. After that, we use the mesh-LBP [6] as a face feature extractor.

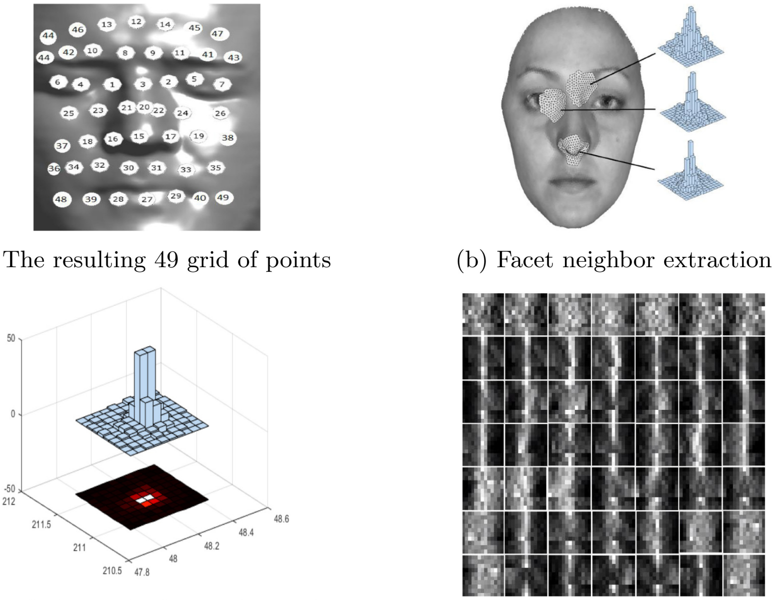
3D Face Modelling: We use the Surrey Face Model [5] for 3D face representation. These open source library provided includes methods to fit the pose and the shape of a model and perform face frontalization. This model is composed of two component: The first component is pose fitting. Given a set of 2D landmark locations and their correspondences in the 3D Morphable Model, the purpose is to estimate the pose of the face.

The second component consists of reconstructing the 3D shape based on the estimated camera matrix. The pose estimation and shape fitting process could be iterated in order to refine the estimates.

Mesh-LBP: The main advantage of the mesh-LBP is the fuse of geometric and appearance features extracted from 3D face models. In the standard LBP (2D-LBP) based face representation [7], we start by dividing the 2D face image into a grid of rectangular blocks, then an histograms of LBP descriptors are extracted from each block and concatenated in order to form a global description of the face.

To extend this workflow to the 3D face model, we need first to split the facial surface into a grid of regions. Then, we compute their corresponding histograms, and group them into a single structure.

The face descriptor construction process is illustrated in the Fig. 1



(a) The resulting 49 grid of points

(b) Facet neighbor extraction

(c) Calculation of the multi-resolution mesh-LBP descriptor

(d) Final face descriptor image

Fig. 1. Face image descriptor construction process

First, the plane formed by the nose tip and the two eyes inner-corner landmark points is initially computed. In fact, the use of only these three landmarks is not arbitrary. But, these points are considered as the most accurate detectable landmarks on the face. Moreover, they are quite robust to facial expressions. Afterwards, the plane is tilted slightly, by a constant amount, to make it more aligned with the face orientation, and then we project this set of points on the face surface, along the plane's normal direction. The outcome of this procedure is an ordered grid of points, which defines an atlas for the facial regions that will divide the facial surface. The grid contain 49 points forming 7 × 7 constellation as shown in Fig. 1a. Once the grid of points has been defined, we extract a neighborhood of facets around each point of the grid. Each neighborhood can be defined by the set of facets confined within a geodesic disc or a sphere, centered at a grid point (Fig. 1b).

3.3 CNN Architecture and Training

We train our CNN in order to classify the face descriptor image created using mesh-LBP. In our work, we deal with a small neural network since we are dealing with images of face descriptors rather than images of faces. The proposed CNN, presented in Fig. 2.

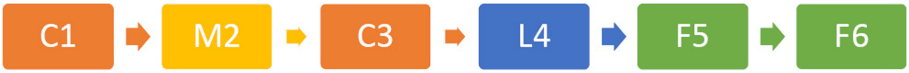


Fig. 2. Architecture of the proposed CNN. The CNN is composed of two convolution layers (denoted by C1, C2), two fully connected layer (F1, F2), max-pooling layer(M) and a locally connected layer (L).

The size of the face descriptor image is 91 × 91 pixels. These images are fed to the our CNN. The first convolutional layer (C1) have 32 filters with size 11 × 11. The resulting 32 feature maps are then fed to a 3 × 3 max-pooling layer (M1) with a stride of 2, separately for each channel. Followed by another convolutional layer (C2) with 16 filters of size 9 × 9. The subsequent layers (L1) is a locally connected layer composed of 16 filter.

Finally, the last two layers, F5 and F6 are fully connected layers. These layers are able to capture correlations between distant face features. The output of the first fully connected layer (F1) in the network is used as our raw face representation feature vector throughout this paper. The output of the last fully-connected layer F2 is fed to a K-way softmax (where K is the number of classes) which produces a distribution over the class labels. It is important to mention the use of the ReLU [32] activation function after the convolution, locally connected and fully connected layer (except the last one L6). In addition, we use the cross-entropy loss in order to maximize the probability of the correct class (face id).

We train our architecture with around 500.000 images from the CASIA-WebFace [33], which contains 494,414 images of 10,575 subjects collected from the Internet. As a first experiment, we are working on face descriptor image, we use a smaller batch size of 200, and we train the network for 10 epochs over the whole data.

4 Experimental Results

In this section, we first present the datasets used in the experiment process. Then, we evaluate our method for face recognition against several challenges, including pose, illumination and face expression variation. Finally, we test our methods in various environments (controlled and crowded)

4.1 Datasets

In this evaluation, we use four datasets:

- **The CMU Multi-PIE face dataset** [8]: It contains more than 750,000 images of 337 people recorded in up to four sessions over the span of five months. Subjects were imaged under 15 viewpoints and 19 illumination conditions while displaying a range of facial expressions.
- **The Bosphorus dataset** [9]: It contains 4666 scans of 105 subjects scanned in different poses, action units, and occlusion conditions, Divided in multiple subsets corresponding to neutral and expressive: Anger, disgust, fear, happy, sad, surprise.
- **The LFW dataset** [10]: It consists of 13,323 web photos of 5,749 celebrities which are divided into 6,000 face pairs in 10 splits.
- **The YTF dataset** [11]: It collects 3,425 YouTube videos of 1,595 subjects (a subset of the celebrities in the LFW). These videos are divided into 5,000 video pairs and 10 splits and used to evaluate the video-level face verification.

4.2 Pose and Illumination-Invariant Face Recognition (PIFR)

The results presented in the Table 1, compares our method against other methods for Pose and illumination-invariant face recognition (PIFR). In other words, we evaluate face recognition while varying illumination and pose.

Table 1. Recognition rate (%) on the Multi-PIE dataset [8] across pose and illumination variations

Method	$-45°$	$-30°$	$-15°$	$+15°$	$+30°$	$+45°$
DNN-CPF [28]	73	81.7	98.4	89.5	80.4	70.3
LNFF-LRA [29]	77.2	87.7	94.9	94.8	88.1	76.4
HPN [30]	71.3	78.8	82.2	86.2	77.8	74.3
U-3DMM [13]	73.1	86.9	93.3	91.3	81.2	69.7
ESO-3DMM [14]	80.8	88.9	96.7	97.6	93.3	81.1
GM-3DMM [31]	84.3	89.4	97.4	99	96.8	92
Proposed Method	**97.4**	**99.5**	**99.5**	**99.7**	**99.0**	**96.7**

The state of the art method could be classified into two subsets. Deep learning based methods [28–30] and 3D based methods [13,14,31]. Our method outperform both deep learning and 3D based methods, and takes benefit from both technologies. By analyzing the results of the 3D based models [13,14,31], we could notice that the use of a 3DMM is well adapted to deal with extreme variations in pose and illumination. Besides, our method obtain much more interesting results, and this is more notable in right and left profile.

To conclude, we notice that the use of the mesh-LBP on 3D data are useful to provide a robust facial feature against illumination and pose.

4.3 Facial Expression Invariant Face Recognition

We tested our method on the Bosphorus dataset, which present seven variation in facial expressions. Results are presented in Table 2.

Table 2. Recognition rate (%) across facial expressions on the Bosphorus dataset

Method	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise
Jingxin et al. [34]	–	71.2	69.3	63.4	90	61	89
Hariri et al. [22]	–	86.25	85.25	81	93	79.75	90.50
Sharma et al. [21]	98.7	94.2	95.7	97.9	96.6	87.3	91.2
Lei et al. [24]	98.96	94.12	88.24	98.55	98.08	96.08	96.92
Deng et al. [23]	100	95.8	92.8	97.7	95.3	98.5	98.6
Abbad et al. [25]	100	95.77	88.41	81.41	88.68	96.97	92.96
Zhang et al. [26]	100	81.69	79.71	88.57	96.23	90.91	95.77
Deng et al. [27]	100	97.2	94.2	97.1	96.2	98.5	98.6
Mesh-LBP [6]	100	97.18	85.51	98.57	88.68	96.97	97.18
Proposed Method	**100**	**97.18**	**96.75**	**100**	**97.63**	**98.88**	**100**

Considering all results, we note that our method is more efficient than the state-of-the-art methods. In addition, the accuracy obtained for neutral emotion is always the highest and several methods achieve 100% accuracy since neutral face is the most common emotion. However, this accuracy decreases while varying facial expressions. Furthermore, disgust and sadness are measured with the lowest accuracy because these emotions are usually unpredictable.

On the one hand, when comparing our method and the Mesh-LBP, we observe that our results and those of the Mesh-LBP are competitive. Furthermore, our method achieves better results, in particular for the DISGUST and HAPPY emotions with an improvement of 10% and 8% respectively. This evolution is due to the learning process and the descriptors extracted from the 3D data

4.4 Face Verification

We evaluate our model against deep face recognition methods on LFW and YTF datasets. Results are presented in the two Tables 3a and 3b. the methods in Table 3a use face image generation for enhancing face recognition results. These methods provide good result, but it is still limited. This limitation is due mainly to the images used in the generation process or the recognition method used.

Table 3. Face verification ratio using the LFW and YTF datasets

Method	Accuracy (%)
LFW-HPEN [12]	96.25
FF-GAN[16]	96.42
DED-GAN [17]	97.52
FI-GAN [35]	98.3
DA-GAN [15]	99.56
Proposed method	**99.59**

(a) Face verification on the LFW dataset.

Method	Accuracy (%)
Deep ID + [37]	93.2
FaceNet [36]	95.12
VGG-face [19]	97.3
Center loss [18]	94.9
DFCL [20]	96.06
Proposed method	**94.97**

(b) Face verification on the YTF dataset.

Looking to Table 3b, our method do not achieve higher values such as [19, 20, 36]. But, improvement is always possible. Our method outperforms some of the well known deep learning method, and provide results that are concurrent to other methods.

5 Conclusion

Face recognition is considered as one of the most complex systems in the field of pattern recognition due to many constraints that are cased by face image appearance variation (accessories, occlusion, illumination, resolution).

In this paper, we propose to combine a 3D model-based alignment, an LBP descriptor constructed on the 3D mesh with a CNN model in order to predict facial identity. The obtained results are quite convincing. Thus, we could conclude that our method achieve higher rates compared to state of the art methods. While indicating that our method does not surpass some others. On the basis of the promising findings presented in this paper, work on the remaining issues is continuing and will be presented in future papers.

References

1. Oloyede, M.O., Hancke, G.P., Myburgh, H.C.: A review on face recognition systems: recent approaches and challenges. Multimedia Tools Appl. **79**(37), 27891–27922 (2020)
2. Anwarul, S., Dahiya, S.: A comprehensive review on face recognition methods and factors affecting facial recognition accuracy. In: Proceedings of International Conference on Robotics and Intelligent Control ICRIC, pp. 495–514 (2020)
3. Yin, Y., Jiang, S., Robinson, J.P., Fu, Y.: Dual-attention GAN for large-pose face frontalization. arXiv preprint arXiv:2002.07227 (2020)
4. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res., 1755–1758 (2009)
5. Huber, P., et al.: A multiresolution 3D morphable face model and fitting framework. In: Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 79–86 (2016)

6. Werghi, N., Tortorici, C., Berretti, S., Del Bimbo, A.: Boosting 3D LBP-based face recognition by fusing shape and texture descriptors on the mesh. IEEE Trans. Inf. Forensics Secur. **11**(5), 964–979 (2016)
7. Wang, H., Hu, J., Deng, J.: Face feature extraction: a complete review. IEEE Access, 6001–6039 (2018)
8. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image Vis. Comput. **28**(5), 807–813 (2010)
9. Savran, A., et al.: Bosphorus database for 3D face analysis. In: Schouten, B., Juul, N.C., Drygajlo, A., Tistarelli, M. (eds.) BioID 2008. LNCS, vol. 5372, pp. 47–56. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89991-4_6
10. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on Faces in Real-Life Images: Detection, Alignment, and Recognition (2008)
11. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: IEEE CVPR, pp. 529–534 (2011)
12. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: High-fidelity pose and expression normalization for face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 787–796 (2015)
13. Hu, G., et al.: Face recognition using a unified 3D morphable model. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 73–89. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_5
14. Hu, G., et al.: Efficient 3D morphable face model fitting. Pattern Recognit. **67**, 366–379 (2017)
15. Yu, Y., Songyao, J., Joseph, P.R., Yun, F.: Dual-attention GAN for large-pose face frontalization. arXiv preprint arXiv:2002.07227 (2020)
16. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Towards large-pose face frontalization in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3990–3999 (2017)
17. Hu, C., Feng, Z., Wu, X., Kittler, J.: Dual encoder-decoder based generative adversarial networks for disentangled facial representation learning. IEEE Access **8**, 130159–130171 (2020)
18. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
19. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015)
20. Deng, W., Chen, B., Fang, Y., Hu, J.: Deep correlation feature learning for face verification in the wild. IEEE Signal Process. Lett. **24**(12), 1877–1881 (2017)
21. Sharma, S., Vijay, K.: Voxel-based 3D face reconstruction and its application to face recognition using sequential deep learning. Multimedia Tools Appl. (1–28) (2020)
22. Hariri, W., Tabia, H., Farah, N., Benouareth, A., Declercq, D.: 3D facial expression recognition using kernel methods on Riemannian manifold. Eng. Appl. Artif. Intell. **64**, 25–32 (2017)
23. Deng, X., Da, F., Shao, H.: Efficient 3D face recognition using local covariance descriptor and Riemannian kernel sparse coding. Comput. Electr. Eng. **62**, 81–91 (2017)
24. Lei, Y., Guo, Y., Hayat, M., Bennamoun, M., Zhou, X.: A two-phase weighted collaborative representation for 3D partial face recognition with single sample. Pattern Recognit. **52**, 218–237 (2016)

25. Abbad, A., Abbad, K., Tairi, H.: 3D face recognition: multi-scale strategy based on geometric and local descriptors. Comput. Electr. Eng. **70**, 525–537 (2018)
26. Zhang, Z., Da, F., Yu, Y.: Data-free point cloud network for 3D face recognition. arXiv, arXiv-1911 (2019)
27. Deng, X., Da, F., Shao, H., Jiang, Y.A.: Multi-scale three-dimensional face recognition approach with sparse representation-based classifier and fusion of local covariance descriptors. Comput. Electr. Eng. **85** (2020)
28. Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., Kim, J.: Rotating your face using multi-task deep neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 676–684 (2015)
29. Deng, W., Hu, J., Wu, Z., Guo, J.: Lighting-aware face frontalization for unconstrained face recognition. Pattern Recognit. **68**, 260–271 (2017)
30. Ding, C., Tao, D.: Pose-invariant face recognition with homography-based normalization. Pattern Recognit. **66**, 144–152 (2017)
31. Koppen P, et al.: Gaussian mixture 3D morphable face model. Pattern Recognit. **74**, 617–628 (2018)
32. Dahl, G.E., Sainath, T.N., Hinton, G.E.: Improving deep neural networks for LVCSR using rectified linear units and dropout. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8609–8613 (2013)
33. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
34. Jingxin, B., Yinan, L., Shuo, Z.: 3D multi-poses face expression recognition based on action units. In: International Conference on Information Technology and Computer Communications (2019)
35. Rong, C., Xingming, Z., Yubei, L.: Feature-improving generative adversarial network for face frontalization. IEEE Access **8**, 68842–68851 (2020)
36. Schroff, F., Dmitry, K., James, P.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
37. Taigman, M.L.Y., Yang, M.: Deep learning face representation from predicting 10,000 classes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1891–1898 (2014)
38. Guo, G., Na, Z.: A survey on deep learning based face recognition. Comput. Vis. Image Underst. **189**, 102805 (2019)
39. Gui, J., Sun, Z., Wen, Y., Tao, D., Ye, J.: A review on generative adversarial networks: algorithms, theory, and applications. arXiv preprint arXiv:2001.06937 (2020)
40. Ning, X., Nan, F., Xu, S., Yu, L., Zhang, L.: Multi-view frontal face image generation: a survey. Concur. Comput. Pract. Exp. (2020)