Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 3699 – 3712
© TÜBİTAK
doi:10.3906/elk-1810-192

Enhancing face pose normalization with deep learning

Anıl ÇELİK[1,*], **Nafiz ARICA**[2]
[1]Cognitus, Etiya, İstanbul, Turkey
[2]Department of Computer Engineering, Faculty of Engineering, Bahçeşehir University, İstanbul, Turkey

Received: 26.10.2018 • **Accepted/Published Online:** 08.04.2019 • **Final Version:** 18.09.2019

Abstract: In this study, we propose a hybrid method for face pose normalization, which combines the 3-D model-based method with stacked denoising autoencoder (SDAE) deep network. Instead of applying a mirroring operation for the invisible face parts of the posed image, SDAE learns how to fill in those regions by a large set of training samples. In the performance evaluation, we compare the proposed method to four different pose normalization methods and investigate their effects on facial emotion recognition and verification problems in addition to visual quality tests. Methods evaluated in the experiments include 2-D alignment, 3-D model-based method, pure SDAE-based method, and generative adversarial network-based normalization method. Experiments performed on Multi-PIE dataset show that the proposed method produces visually reasonable results and outperforms the others in facial emotion recognition. On the other hand 2-D alignment is sufficient in the verification problem where the detailed face characteristics should be preserved in the normalization process.

Key words: Face frontalization, pose normalization, deep learning

1. Introduction

Nowadays, computer vision research on face related problems are gaining considerable popularity due to increasing number of facial applications such as face recognition, emotion analysis. Unfortunately, face images used in such applications are not always acquired in, so to say, ‘ideal‘ conditions. Pose variances that faces may happen to have could make the faces appear drastically different than their poseless counterparts; certain parts of the face may be less visible, or not visible at all. Therefore, it is crucial to frontalize the face images by some set of pose normalization procedures before the subsequent processes. As the pose normalization directly affects performance of the applications, research on solving this problem is gaining popularity as well.

Face pose normalization approaches in the literature can be generally classified under 2-D and 3-D methods. 2-D methods can be categorized under three different groups inside themselves: piece-wise warping, patch-wise warping, and pixel-wise warping. Piece-wise warping approaches transform the shape of the face image in a piecewise manner to another specified pose. Piece in this approach refers to a single triangle in the triangular mesh, generated by triangulation of the facial feature points. Piece-wise–based methods aim to warp the target image mesh to the original mesh using a geometric transformation. The aforementioned transformations include, but are not limited to, affine warping [1] and thin-plate splines-based warping [2]. Patch-wise warping methods, on the other hand, are known for their ability to cover a wider variety of poses, which is one of the main problems in piece-wise warping methods. One of the recent and notable works

[*]Correspondence: anilcelik@yahoo.co.uk

models the face image as a collection of patches and accomplishes the reconstruction of face images using a patch-wise strategy, calling it the 'stack-flow' approach [3]. Taking the strategy one step further, the approach in [4] takes overlapping pixels into consideration and proposes an optimal set of local warps for face pose normalization. To alleviate the problem on poor handling of local nonlinear warps that appear in each piece or patch-based approaches, the authors in [5] propose the spearheading parallel-deformation method, which predicts the pixel-wise displacement between two poses. The downside of the approach in [5] is establishing pixel-wise correspondence between images. Other examples of pixel-wise approach can be found in [6] and [7], which propose template displacement fields generated from a set of 3-D face models for face pose normalization.

The human head is a complex 3-D structure, rotating in 3-D space while the face image lies in the 2-D domain makes it difficult to conduct face synthesis with the limitations of 2-D techniques. As an alternative to 2-D, 3-D–based methods build a model of the human head to conduct frontal posed face creation. The 3-D pose normalization approaches employ the 3-D facial shape model as a tool to correct the nonlinear warp of facial textures appearing in the 2-D images. The general principle is that, 2-D face image is first aligned with a 3-D face model, typically with the help of facial landmarks. Then, texture of the 2-D image is mapped to the 3-D model. Lastly, textured 3-D model is rotated to a desired pose and a new 2-D image in that pose is rendered. Early approaches utilize simple 3-D models, like cylinder model [8], wireframe model [9], and ellipsoid model [10] to roughly model the 3-D structure of the human head, whereas newer approaches strive to build accurate 3-D facial shape models. A recent notable example is Tal Hassner's novel approach [11], which can be considered one of the most influential works in the field. Using a generic 3-D reference model, the approach estimates the camera matrix used to capture query image. Then, the aforementioned matrix is processed to relocate pixels to create the initial pose normalized face. Finally, the seminormalized face is fine-tuned with a heuristic mirroring method to fill missing regions due to the geometric transformation. Inspired by [11], the approach in [12] used a very similar method; however, this time instead of a generic 3-D face model, a morphable face model is employed, namely, the Basel Face Model[1], which allows the approach to push the 3-D methodology even further by allowing effective control over emotions. In this case, the authors chose to normalize the expression with pose, which results in truly neutral, poseless faces.

Currently, deep learning is becoming very popular in many different fields. It would be unrealistic to think that it would not be applied in face pose normalization. For example, the authors in [13] employ a deep network for this specific task. Defining the undesired pose angles as a type of noise, a SDAE is employed to effectively normalize faces. At each iteration group, the weights on the finalizing layers are saved, to be passed down to a new, deeper network. This way, valuable weight data is preserved. Resulting in a pose normalization framework, which starts shallow, then gradually expands into a deep network. Surely, autoencoders are not the only deep learning approach that is viable; convolutional neural networks (CNN) also get a respectable amount of exposure. A portion of its popularity is gained with the discovery of the fact that CNN's powerful feature extraction capability could be used in a different manner. Starting with the observation that facial feature vectors extracted from different poses of the same person can still produce the same feature vector, the approach in [14] uses a CNN architecture to extract face features from posed faces in such a manner that they would be reshaped into frontal ones as they passed through the network. Finally, a new architecture called generative adversarial network (GAN) has been introduced to the field of deep learning recently, and the method in [15] is directly involved with pose normalization using GAN. This study brings three novelties to the table;

[1] faces.cs.unibas.ch. 2009. Basel Face Model (BFM) [online] available at: http://faces.cs.unibas.ch/bfm [Accessed 22 Aug. 2015]

firstly, encoder-decoder structure of the generator enables the network to learn a representation that is both generative and discriminative, which, in the end, can be employed for pose normalization and pose-invariant face recognition. Secondly, this representation is explicitly disentangled from other face variations such as pose, through the pose code provided to the decoder and pose estimation in the discriminator. Finally, the method in [16] can take one or multiple images as the input, and generate one unified identity representation along with an arbitrary number of synthetic face images.

As a result, all of the approaches available in the literature have some pros and cons, and face pose normalization problem is not fully solved yet. In this study, we aim to combine two recent approaches to improve the performance. For this purpose, we propose to enhance 3-D pose normalization by deep learning. In the experiments, we compare the proposed method with four different approaches. The comparisons start from 2-D alignment, then get gradually more sophisticated, to SDAE network-based pose normalization, 3-D model-based pose normalization, and GAN-based pose normalization. 2-D alignment method applies an affine transformation to faces, based on the facial feature points. It is simple and cost-efficient. In deep learning method, we design a SDAE network, which defines the face pose variations as a type of noise. 3-D model-based approach employs the method in [11], which normalizes posed faces by estimating the camera matrix used to capture the query image. It benefits from the correspondence between facial landmark points on query image and 3-D reference model. The GAN-based method [15] learns a disentangled representation by performing face frontalization and pose-invariant representation together. Our proposed approach enhances the 3-D model-based method by adding SDAE network as a postprocessing tool to improve the final frontal face.

In performance analysis, all of the approaches are evaluated under three types of experiments. The first one is the visual quality, by checking on the appearance of resulting faces, to see if they preserve identity and facial expression. The other problems are facial emotion recognition and face verification. In the experiments, we use Multi-PIE dataset[2], which includes over 750.000+ images taken from 300+ unique people. Different expressions, pose angles, and lighting conditions are included for each person. This collection of different conditions and especially, variety in poses is invaluable. In visual inspection experiments, it is seen that combining SDAE with 3-D model–based normalization improves both pure SDAE-based and 3-D model–based methods. While 2-D alignment does not change the visual quality of posed input image, GAN-based frontalization distorts the facial characteristics for the sake of producing relatively higher-quality face images. In face verification experiments, 2-D alignment verifies the face identities better than all the others. It is due to the fact that 2-D alignment does not cause any distortion even in a pixel. Finally, facial emotion recognition experiments show that the proposed method produces significantly higher recognition rates than all the other methods. As a result of all the experiments performed on Multi-PIE dataset, we can say that 2-D alignment might be sufficient in the applications where detailed face characteristics should be preserved in normalization process. On the other hand, our proposed pose normalization produces better results in the applications requiring the discrimination of general facial characteristics.

The rest of the paper is organized as follows; in Section 2, after giving an overview of the proposed method, we explain each procedure in detail in the subsequent sections. Section 5 gives experimental results in visual quality, emotion recognition, and face verification. Finally, Section 6 concludes the study and briefly gives ideas about future work.

[2] CMU Multi-PIE Face Database. [online] Available at: http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/MultiPie/Home.html [Accessed 10 Apr. 2016]

2. System overview of the proposed method

The proposed study mainly employs three consecutive stages; preprocessing, 3-D pose normalization, and SDAE-based fine-tuning. Given a posed face image, the first step is to apply preprocessing operations, starting with landmark detection. Then, those facial landmark points are used to 2-D align faces. To finish off, a lighting normalization process is employed. Next step is to feed 2-D aligned images into 3-D pose normalization stage, which consists of pose estimation and frontal pose generation. After face images go through semipose normalization, they are fed into a deep network with their poseless counterparts. Finally, the SDAE deep network finalizes the pose normalization operation by fine-tuning the resulting image of 3-D model-based method. The proposed study is visualized in Figure 1.

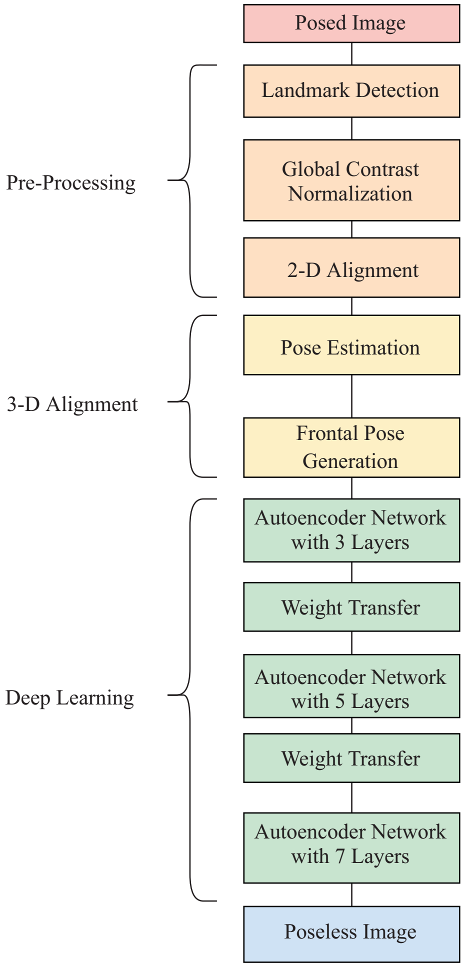

Figure 1. Overview of the proposed approach.

3. Preprocessing

3.1. Landmark detection

Landmark detection is accomplished with the approach in [16]. The algorithm estimates landmark positions by employing an ensemble of regression trees. Implementation of the algorithm can be found in the software library, Dlib[3], trained with the images from HELEN[4] dataset. Unlike the original paper in [16], we detect 68 unique facial feature points to make it compatible with 3-D face model in [11]. An example of landmark detection result is given below in Figure 2. In the training stage, we use Multi-PIE dataset containing 3 unique angles of poses (15°, 30°, and 45°).

3.2. Global contrast normalization

Global contrast normalization (GCN) is a preprocessing step that prevents images from having varying degrees of contrast, by subtracting the mean from each image and then rescaling, so that the standard deviation across its pixels is equal to a constant. An example result of the process can be examined below in Figure 3.

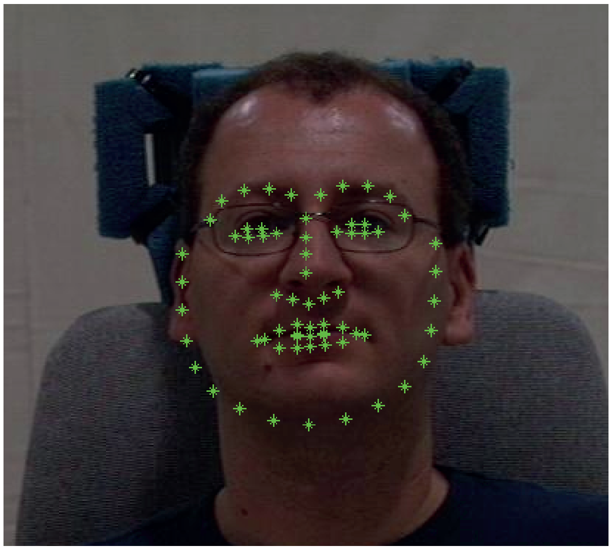

Figure 2. Sixty-eight landmark points located with the approach in [16].

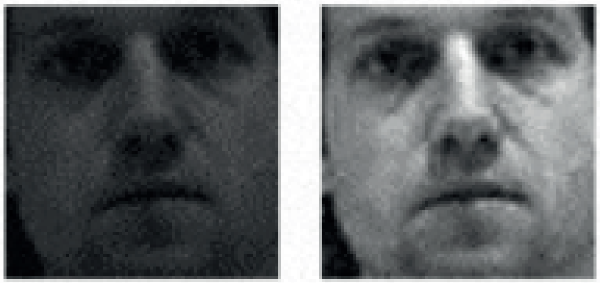

Figure 3. Sixty-eight original image (left), image with GCN applied (right).

3.3. 2-D alignment

2-D alignment is employed to align facial features, so that deep networks have an easier time learning features. Performed by affine transformation, given a point pair p and q and matrices A and B, transformation can calculate a new set of points x and y.

$$\begin{bmatrix} x \\ y \end{bmatrix} = A \times \begin{bmatrix} p \\ q \end{bmatrix} + B \tag{1}$$

In our approach, [x,y] and [p,q] point pairs represent the landmarks captured from posed and frontal face images. Leaving the spatial transformation matrix to be calculated, and then applied to images for 2-D alignment. An example 2-D alignment can be examined below in Figure 4.

[3] Dlib.net. 2016. dlib C++ Library. [online] Available at: http://dlib.net [Accessed 19 Dec. 2016]

[4] ifp.illinois.edu. 2012. HELEN Dataset. [online] Available at: http://www.ifp.illinois.edu/ vuongle2/helen/ [Accessed 2 Apr. 2017]

4. 3-D pose normalization enhanced by deep learning

The motivation behind this study is to improve the 3-D pose normalization results by adding some learning procedure to estimate invisible parts in frontal pose generation. 3-D pose normalization process in [11] is the backbone of our approach. It is a viable pose normalization approach. Pose normalization in [11] is defined as the process of back-projecting the appearance of the query face image to the reference coordinate system using the 3-D surface as a proxy. The approach is developed using a nonpersonalized, preestablished 3-D model to approximate the camera matrix. Based on the camera matrix, query face is reshaped into a frontal one, using bilinear interpolation.

Pose normalization in [11] originally uses five main steps; namely, landmark detection, pose estimation, frontal pose generation, visibility estimation, and mirroring. Facial landmarks detected in 2-D alignment is used in the first step. Pose estimation and frontal pose generation are explained in the following subsections. In this study, we do not use visibility estimation and mirroring; instead, we employ deep learning procedure to improve the frontalization performance.

4.1. Pose estimation

Using the reference 3-D model, a 2-D rendered version of it is generated with a projection matrix. A frontal version of the 3-D reference model is synthesized to serve as a reference view. During its synthesis, 3-D coordinates of each pixel are stored. Then, 2-D landmarks located in the query image are located in the reference view. Exploiting the correlation between landmarks in query image and 3-D landmarks in reference model, camera matrix used to capture query image is estimated.

4.2. Frontal pose generation (soft pose normalization)

In this preprocessing step, pixels of query image are back-projected, creating frontal posed face image. This part of the approach in [11] generates the input of deep learning enhanced pose normalization. The result of this step can be examined below in Figure 5.

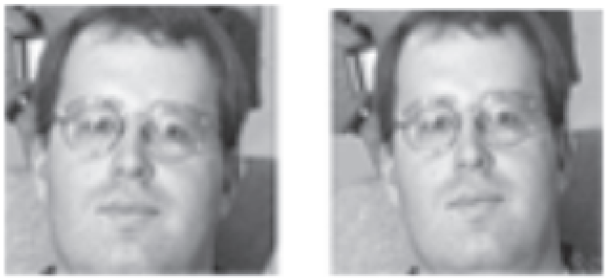

Figure 4. Original image (left), 2-D aligned image (right).

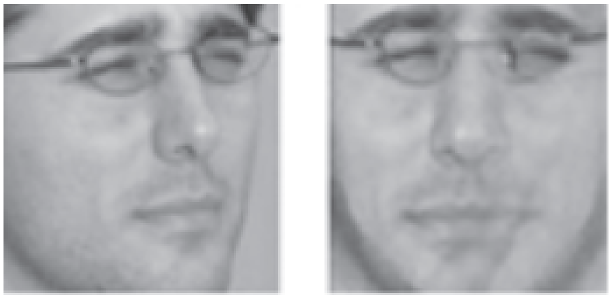

Figure 5. Original image (left), soft pose normalized image (right).

4.3. Enhancing 3-D pose normalization using SDAE

The motivation behind this study is to improve the 3-D pose normalization results by adding some learning procedure to estimate the invisible parts in frontal pose generation. The approach in [11] constructs a visibility map on the seminormalized face by counting overlapping projected pixels, due to increase in the angle between the face and the camera. An example visibility map is visualized in Figures 6a–6c.

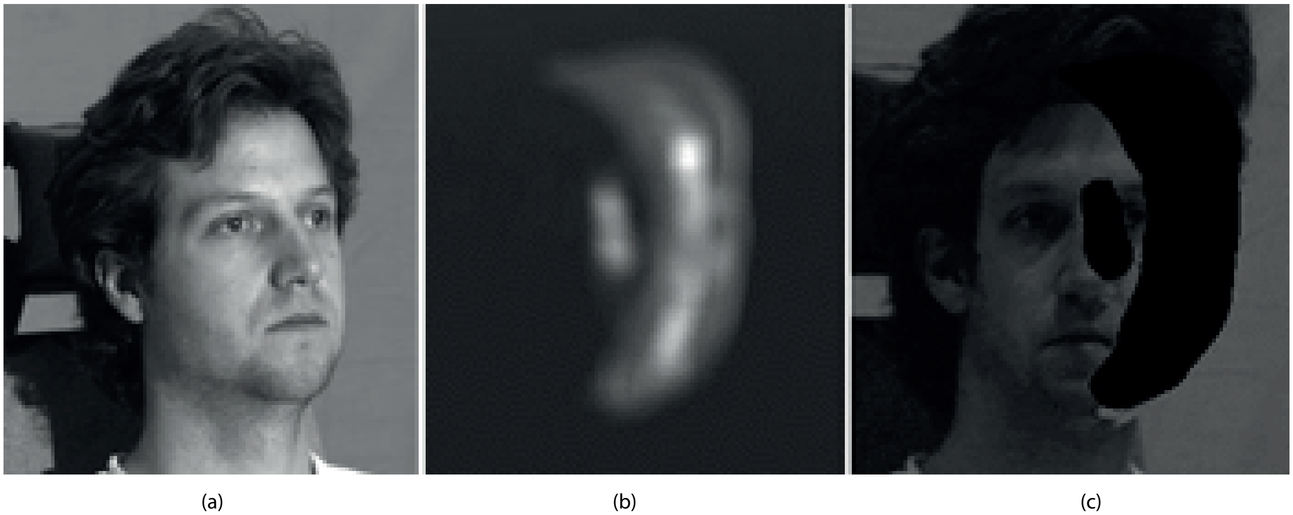

Figure 6. (a) Original image, (b) visibility map, (c) visibility map projected over soft pose normalized image.

The next step is to take the mirror of the visible side with the help of the visibility map. Invisible parts in seminormalized face are filled by the mirror of visible parts on the other side to generate a more visually satisfying output. However, the mirroring step is heuristic, meaning that it will apply the same solution to every type of input it is going to encounter, which, in the end, will result in a subpar outcome at best when confronted by a particularly challenging example.

In this study, we employ SDAE to improve 3-D pose normalization on the assumption that the unnatural looking regions in the soft pose normalized outputs of [11] can be interpreted as a type of noise. Input posed images go through the 3-D pose normalization process, but instead of applying a mirroring operation, SDAE is used to estimate the invisible parts. The seminormalized image is fed into the SDAE to finalize the pose normalization process. Workflow of deep learning enhanced 3-D pose normalization is shown below in Figure 7.

The SDAE network starts with 3 layers, and gets progressively deeper after finishing each training cycle, reaching up to 5 and then finally, 7 layers. After each training cycle, weights of layers on the right side of the middle hidden layer are transferred to a larger network. Meaning that, after the first iteration 1, and after the second iteration 2 layers worth of weights will be transferred to the deeper network. This method is similar to transfer learning but not quite the same. Weights of layers closer to the output layer hold valuable information that can be improved by changing the architecture gradually. The methodology behind this design is shown below in Figure 8.

Deep network employs various methods, with the balance of performance and accuracy on focus. Adadelta [17] is employed as optimizer while binary-crossentropy is used as a loss function. For the activation functions, sigmoid is used for the final layer, while the rest consists of rectified linear unit functions (ReLu). ReLu is chosen because of the efficiency due to the lack of exponential computation, since the way it works makes taking derivatives much less taxing. Every training cycle takes 50 epochs to finish. Batch size is 25. The number is chosen as such to let deep network see patterns easily; lowering it would risk a poor fit. Training, test, and validation data split is roughly 70%, 15%, and 15%. The net amount of images used for all is 13,000. The number is kept at this amount due to two factors: computational tax and images not being fit for training due to reasons such as the approach in [11] producing poor pose normalization or faulty 2-D alignment.

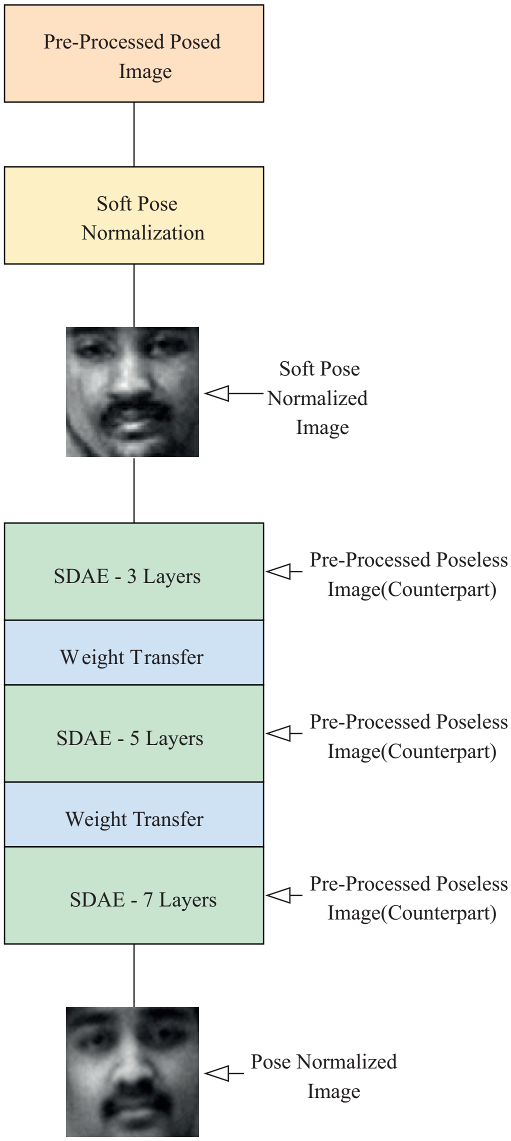
Pre-Processed Posed Image
Soft Pose Normalization
Soft Pose Normalized Image
SDAE - 3 Layers
Pre-Processed Poseless Image(Counterpart)
Weight Transfer
SDAE - 5 Layers
Pre-Processed Poseless Image(Counterpart)
Weight Transfer
SDAE - 7 Layers
Pre-Processed Poseless Image(Counterpart)
Pose Normalized Image

Figure 7. Workflow of enhancing 3-D model-based pose normalization by SDAE.

5. Experiments

In the performance analysis, we compare our method to four different methods; 2-D alignment, 3-D pose normalization in [11], SDAE-based normalization, and finally, GAN-based approach in [15] as an external source, to increase diversity and the validity of the experiments. 2-D alignment is performed after the landmark detection and global contrast normalization as described in the previous sections. In 3-D pose normalization, the original work of [11] is used as it is. After frontal pose generation process, mirroring is applied based on the visibility map. Details of 3-D pose normalization can be found in [11]. The SDAE-based normalization method uses the same architecture in our proposed study. The number of examples is increased to 20,000 with the same split as the previous method. Increase in the number of images is due to the problem of becoming more sophisticated. As expected, posed faces provide a larger variety. This time, instead of the 3-D pose-normalized inputs, 2-D–aligned posed images enter the SDAE network, paired with their poseless counterparts.

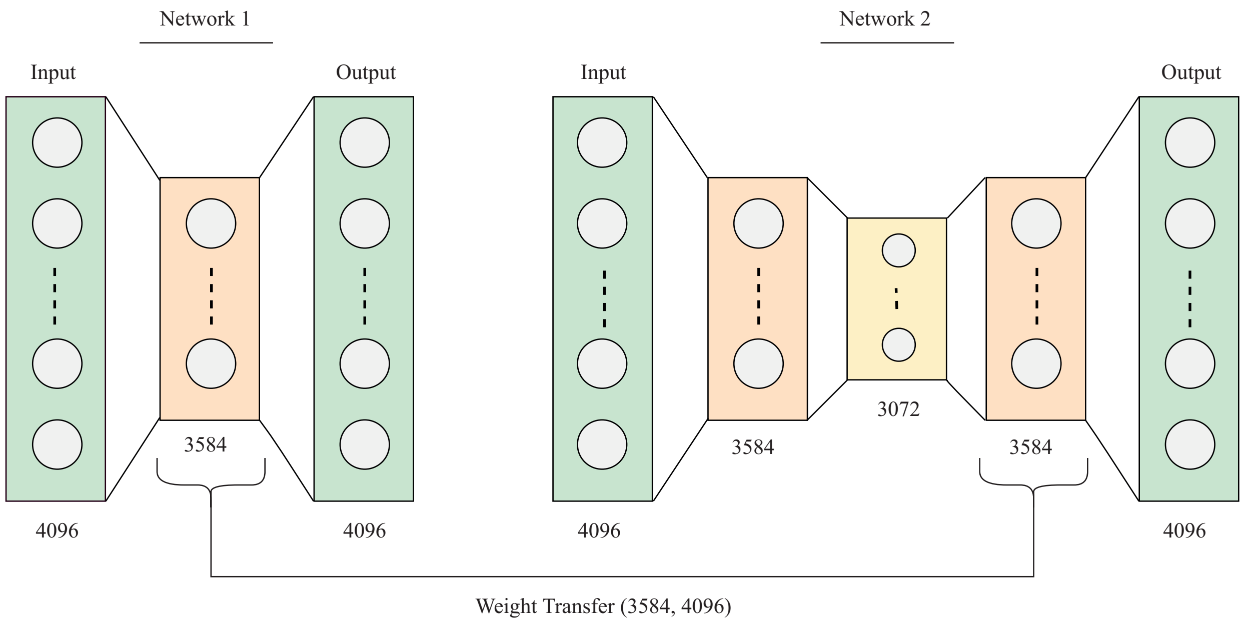

Figure 8. Deep network weight transfer process on our approach.

The method in [15] is tested with the pretrained model provided by the authors. The images are aligned with the method, again provided by the authors. Experiment results are grouped under three sets; visual inspection, face verification, and emotion recognition.

5.1. Setup

Software implementation of the proposed study is realized with MATLAB and Python, working on 2013b version with Visual Studio 2010 Compiler for Mex files and Python 2.7. Multi-PIE is the dataset used in both training and testing. There are three different angles, selected on the left side of the face, being; 15°, 30°, and 45°. For each person in the dataset, 20 different lighting conditions are employed in both training and testing.

Face verification and emotion recognition are evaluated with Microsoft's cloud platform, Azure[5], which employs state-of-the art technology on both fronts.

5.1.1. Face verification method

Azure face verification approach [18] argues that the famous 'curse of dimensionality' can be bypassed in certain cases to gain considerable benefit. Of course, as good as it sounds, high dimensional feature vectors are not used as is, instead features are made practical with a sparse projection approach. Features get compressed, then a sparse projection matrix encodes features to a lower dimension, with the help of regularization. Between 1000 and 100,000 dimensional features, there is almost 7% accuracy improvement in face verification. To extract features, multiple landmark points are located, including eyes, mouth, and nose. Each landmark point patches are extracted in multiple scales. After each patch is divided into a grid, a different descriptor extracts features from their respective subpatch in the grid. This approach being multiscale in this context means that face will get smaller in each iteration; in addition, number of landmarks collected in each iteration will increase. As the

[5] azure.microsoft.com. 2014. Microsoft Azure [online] Available at: https://azure.microsoft.com/en-us/services/cognitive-services/emotion/ [Accessed 23 May. 2017]

face region gets smaller, extracted patch will cover a larger area. This effect is exploitable due to fact that the features will benefit from global shape information. During the verification process, five different feature types are extracted from grid patches: LBP [19], SIFT [20], HOG [21], Gabor [22], and LE [23], then all of features are concatenated into a single vector. Finally, face verification is done with a joint Bayesian method [24] due to the higher accuracy numbers compared to other methods evaluated in the approach.

5.1.2. Emotion recognition method

Capable of recognizing 7 different emotions (angry, disgust, fear, happy, neutral, sad, surprise), Azure emotion recognition API [25] is equipped with 3 state-of-the-art face detection approaches, including Joint cascade detection (JDA) [26], deep convolutional neural network (DCNN) [27] and mixtures of trees (MoT) [28]. Face detectors work as an ensemble, but in a hierarchical way, meaning that the latter will only activate if JDA fails to detect a face in the image. Each detector is proficient in one environmental setting, namely, JDA is highly accurate with faces that are poseless or are close to being poseless. DCNN on the other hand shows impressive results in profile face images.

In emotion recognition task, once again, an ensemble approach is employed, but not in a hierarchical way this time. Ensembling helps acquiring diverse models for the task. In the ensemble, CNNs are randomly initialized. A voting scheme is deployed with weighted voting, to benefit from robust sides of the models in the ensemble.

5.2. Visual inspection

Although visual results do not directly shed light on the performances of recognition and verification, we think that the inspection of visual quality is important to analyze the effects of normalization processes. This section is about inspecting visual results to determine the changes which the face images go through as the methods are applied. The aforementioned results can be examined below in Figures 9a–9e.

Change in normalized faces that 2-D alignment brings is minimal, yet it is sufficiently noticeable. 3-D model-based approach in [11] provides reasonable visual results with the quality close to the original. However, there are synthetic regions over the face. Severity of this effect changes depending on the degree of the pose. These regions happen due to the heuristic mirroring approach employed in the method which fails locate invisible pixels correctly. Unfortunately, this problem diminishes the reality factor in face appearances. Identity of faces are preserved while expressions suffer a hit. The results of pure deep learning by SDAE suffer from the blurring effect. Identity and expression preservation are slightly improved compared to [11]. The proposed approach which enhances 3-D model using SDAE provides better visuals, as expected. Identities and expressions are preserved and there is no visible deterioration in the faces. However, due to the low resolution of the images used in the deep learning step, there is a noticeable blurring. The blurring effect is slightly lower than in the pure SDAE method. Deep learning results also suffer from the same blurring effect. Identity and expression preservation are slightly better. The GAN-based method [15] provides results that are a mixed bag. Faces look realistic. However, at times the network takes 'artistic freedom'. Hair or beard is added to the images, when these do not exist in the original image. While not particularly visible in gray-scale, some faces appear to have a stitch mark in the middle.

5.3. Emotion recognition

Emotion recognition evaluations are employed under 4 different settings (15°, 30°, 45°, mixture of all). Mixture setup takes all the samples from 15°, 30°, and 45° posed samples in the database. In the experiments training

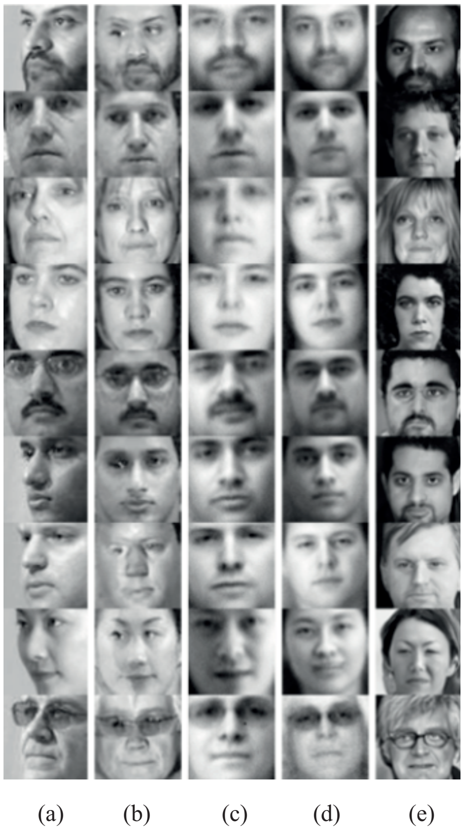

Figure 9. (a) 2-D–aligned posed image, (b) 3-D model-based method [11], (c) SDAE-based method, (d) [11] enhanced by SDAE, (e) GAN-based method [15].

and testing sets are generated separately in each setting. In that sense, mixture setup can be considered as the most realistic experiment. Accuracy measure is determined as the amount of correct recognition results out of all examples evaluated.

Table 1. Emotion recognition results (in percentage).

Pose	2-D alignment	[11]	[11]+deep learning	Deep learning	[16]
45	55	75	87	90	54
30	62	77	90	85	67
15	66	83	85	86	69
Mix.	59	80	89	83	61

While a computationally cheap operation to realize, 2-D alignment provides mediocre results at best when it comes to emotion recognition. Normally, the method in [11] provides lower performance as the problem gets more difficult. Starting at 83% on 15°, it decreases until 75% at 45°. The SDAE-based method provides relatively higher accuracy scores compared to [11], minus the restrictive preprocessing requirements, in addition, operation is relatively less costly. Surprisingly, pure deep learning by SDAE provides higher performances as the pose angle increases. The GAN-based method [15] is closer to 2-D alignment in terms of performance. Although image quality seems better than the other methods, it has the similar performances as in 2-D alignment. If the accuracy is important above all, the proposed method provides the best performances in the mixture setup. It has also the best accuracy point at 30°. We think that having the best accuracy at the middle pose angle is due to [11]'s outputs. Around 30°, 3-D model-based method visually preserves the emotions best. Another advantage of our method and the pure SDAE-based method is homogeneous accuracy values across the different pose angles. In both methods, the recognition rates are between 85% and 90%. Deep learning by itself provides accuracy numbers that can compete with those of the approach in [11], minus the restrictive preprocessing requirements; in addition, operation is relatively less costly.

5.4. Face verification

Face verification evaluations are employed under the exact same settings used in emotion recognition evaluations.

Table 2. Face verification results (in percentage).

Pose	2-D alignment	[11]	[11]+deep learning	Deep learning	[16]
45	73	53	66	59	54
30	78	55	64	65	61
15	86	60	67	62	66
Mix.	81	58	71	70	63

2-D alignment approach yields the highest performance in this experiment. It is cost-effective and the most accurate of all the methods tested. It is affected by the increasing pose degrees negatively. In one way or another, all of the other approaches change the structure of the face. Face verification used in the experiments is very sensitive to such changes. In [11], the heuristic mirroring approach makes the accuracy numbers suffer hits across the board. Large pose values make the corruptions grow exponentially, thus damaging the identity of the face. Thus, [11] has the lowest scores among the other methods. The GAN-based method [15] provides results close to [11] due to the corrections in face images. The SDAE-based method provides higher performances than both [11] and [15], by having verification accuracies between 59% and 70%. The proposed method is slightly better than those three methods which distort the face images for the sake of producing visual quality results.

6. Conclusion

In this study, we propose a hybrid method for face frontalization problem. It enhances the 3-D model based method by a SDAE deep network. We also investigate the effects of four different pose normalization methods on different face based computer vision applications including facial emotion recognition and verification. While our hybrid deep learning approach yields high performance in emotion recognition, it falls behind slightly on face verification due to the loss of identity. 2-D alignment produces the best results on face verification, since the method is all about minimizing operations done over the face region thus preserving identity well. SDAE deep network alone provides very close results to our proposed method on both cases and proves itself to be a worthy alternative for those who want a balanced solution. 3-D model–based pose normalization gives the middle-of-the-pack results overall, and stays as an option worth taking in emotion recognition. Evaluations show that SDAE deep network can improve the accuracy of 3-D normalization in certain conditions; however, they also show that there are still shortcomings that come with the methodology. In the future, we aim to push the deep learning-based approaches further to improve the face verification accuracy. The promise shown by the approach in expression recognition tells us that there is much more to be exploited in the deep learning field, to provide unexpected solutions.

Acknowledgment

This work was developed under grant '115E310' from the Scientific and Technological Research Council of Turkey (TÜBİTAK).

References

[1] Zhang X, Gao Y. Face recognition across pose: A review. Pattern Recognition 2009; 42 (11): 2876-2896. doi: 10.1016/j.patcog.2009.04.017

[2] Bookstein F. Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence 1989; 11 (6): 567-585 doi: 10.1109/34.24792

[3] Ashraf A, Lucey S, Chen T. Learning patch correspondences for improved viewpoint invariant face recognition. IEEE Conference on Computer Vision and Pattern Recognition 2008; Anchorage, Alaska, USA; 2008. pp.1-7. doi: 10.1109/CVPR.2008.4587754

[4] Ho H, Chellappa R. Pose-invariant face recognition using Markov random fields. IEEE Transactions on Image Processing 2013; 22 (4): 1573-1584. doi: 10.1109/TIP.2012.2233489

[5] Beymer D, Poggio T. Face recognition from one example view. Proceedings of IEEE International Conference on Computer Vision 1995; Cambridge, MA, USA; 1995. pp. 500-507. doi: 10.1109/ICCV.1995.466898

[6] Li S, Liu X, Chai X, Zhang H, Lao S et al. Morphable displacement field-based image matching for face recognition across pose. European Conference on Computer Vision 2012; Firenze, Italy; 2012. pp. 102-115. doi: $10.1007/978-3-642-33718\_8$

[7] Li S, Liu X, Chai X, Zhang H, Lao S et al. Maximal likelihood correspondence estimation for face recognition across pose. IEEE Transactions on Image Processing 2014; 23 (10): 4587-4600. doi: 10.1109/TIP.2014.2351265

[8] Gao Y, Leung M, Wang W, Hui S. Fast face identification under varying pose from a single 2-D model view. IEE Proceedings - Vision Image and Signal Processing 2001; 148 (4): 248-253. doi: $10.1049/ip-vis{:}20010377$

[9] Lee M, Ranganath S. Pose-invariant face recognition using a 3D deformable model. Pattern Recognition 2003; 36 (8): 1835-1846. doi: 10.1016/S0031-3203(03)00008-6

[10] Liu X, Chent T. Pose-robust face recognition using geometry assisted probabilistic modeling. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005; 502-509. doi: 10.1109/CVPR.2005.276

[11] Hassner T, Harel S, Enbar R. Effective face frontalization in unconstrained images. IEEE Conference on Computer Vision and Pattern Recognition 2014; Columbus, Ohio, USA; 2014. pp. 4295-4304. doi: 10.1109/CVPR.2015.7299058

[12] Zhu X, Lei Z, Yan J, Yi D, Li S. High-fidelity pose and expression normalization for face recognition in the wild. IEEE Conference on Computer Vision and Pattern Recognition 2015; Salt Lake City, UT, USA; 2015. pp. 787-796. doi: 10.1109/CVPR.2015.7298679

[13] Kang Y, Lee K, Eun J, Park S, Choi S. Stacked denoising autoencoders for face pose normalization. International Conference on Neural Information Processing 2013; Daegu, South Korea; 2013. pp. 241-248. doi: $10.1007/978-3-642-42051-1\_31$

[14] Zhmoginov A, Sandler M. Inverting face embeddings with convolutional neural networks. Published in ArXiv 2016; arXiv:1606.04189

[15] Tran L, Yin X, Liu X. Disentangled Representation Learning GAN for pose-invariant face recognition. IEEE Conference on Computer Vision and Pattern Recognition 2017; Honolulu, HI; 2017.pp.1-10. doi: 10.1109/CVPR.2017.141

[16] Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees. IEEE Conference on Computer Vision and Pattern Recognition 2014; Columbus, OH; 2014. pp.1867-1874. doi: 10.13140/2.1.1212.2243

[17] Zeiler M. ADADELTA: An Adaptive Learning Rate Method. Published in ArXiv 2012; arXiv:1212.5701

[18] Chen D, Cao X, Wen F, Sun J. Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification. IEEE Conference on Computer Vision and Pattern Recognition 2013; Portland, OR, USA; 2013. pp. 3025-3032. doi: 10.1109/CVPR.2013.389

[19] Ahonen T, Hadir A, Pietikainen M. Face description with local binary patterns: Application to face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 2006; 28 (12): 2037-2041. doi: 10.1109/TPAMI.2006.244

[20] Lowe D. Distinctive image features from Scale-Invariant Keypoints. International Journal of Computer Vision 2004; 60 (2): 91-110. doi: 10.1023/B:VISI.0000029664.99615.94

[21] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005; San Diego, CA, USA; 2005.pp. 886-893. doi: 10.1109/CVPR.2005.177

[22] Liu C, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing 2002; 11 (4): 467-476. doi: 10.1109/TIP.2002.999679

[23] Cao Z, Yin Q, Tang X, Sun J. Face recognition with learning-based descriptor. IEEE Computer Society Conference on Computer Vision and Pattern Recognition; San Francisco, CA; 2010. pp. 2707-2714. doi: 10.1109/CVPR.2010.5539992

[24] Chen D, Cao X, Wang L, Wen F, Sun J. Bayesian face revisited: A joint formulation. European Conference on Computer Vision 2012; Firenze, Italy; 2012. pp.566-579. doi: $10.1007/978-3-642-33712-3\_41$

[25] Yu Z, Zhang C. Image based static facial expression recognition with multiple deep network learning. International Conference on Multimodal Interaction; Seattle, USA; 2015. pp. 435-442. doi: 10.1145/2818346.2830595

[26] Chen D, Ren S, Wei Y, Cao X, Sun J. Joint cascade face detection and alignment. European Conference on Computer Vision; Zürich; 2014. pp. 109-122. doi: $10.1007/978-3-319-10599-4\_8$

[27] Zhang C, Zhang Z. Improving multiview face detection with multi-task deep convolutional neural networks. IEEE Winter Conference on Applications of Computer Vision; USA; 2014. pp. 1036–1041. doi: 10.1109/WACV.2014.6835990

[28] Meila M, Jordan M. Learning with mixtures of trees. The Journal of Machine Learning Research 2000; 1 (2000): 1-48. doi: 10.1162/153244301753344605