Crimson Publishers
Wings to the Research

Mini Review

Robust Representation for Face Recognition in the Wild

Mostafa Farag, Mostafa Abdelrahman, Ahmed El-Barkouky, Eslam Mostafa, James Graham and Aly A Farag*

Fellow, IEEE and IAPR Computer Vision and Image Processing Laboratory, University of Louisville, USA

ISSN: 2640-9739

***[1]Corresponding author:** Aly A Farag, Fellow, IEEE and IAPR Computer Vision and Image Processing Laboratory, University of Louisville, Louisville, KY, USA 40292

Submission: April 07, 2020

Published: June 02, 2021

Volume 2 - Issue 1

How to cite this article: Mostafa Farag, Mostafa Abdelrahman, Ahmed El-Barkouky, Eslam Mostafa, James Graham and Aly A Farag*. Robust Representation for Face Recognition in the Wild. COJ Elec Communicat. 2(1).COJEC.000530.2021. DOI: 10.31031/ COJEC.2021.02.000530

Abstract

This paper describes a generalized representation for face recognition in the wild using still and video imaging, addressing unconstrained recognition under A-PIE effects. The proposed representation uses images and video frames for generating the facial signatures for face recognition. The representation allows for inclusion of all video frames in generating the facial signature and is also amenable to "best frames" selection and "intelligent priors" about facial regions in the image/video. The representation includes dual optimization for signature extraction and signature matching over large gallery manifolds. It is modular and amenable to deployment on distributed systems, smart phones and on the cloud. Results, in a Face Recognition at a Distance (FRAD) setup, for detection of faces under severe occlusion, pose-invariant face detection and recognition, on databases created by high resolution camera (Canon 7D: 18MP) and iPhone 4 low-resolution camera (5MP), give credence to the proposed representation.

Keywords: Face recognition in the wild; Selective Part Model (SPM); Gallery manifolds; Pose-invariance; Illumination models; Deep learning; Feature selection; Matching

Introduction

Research in face recognition deals with problems related to Age, Pose, Illumination and Expression (A-PIE), and seeks approaches that are invariant to these factors. Video images add a temporal aspect to the image acquisition process. Another degree of complexity, above and beyond A-PIE recognition, occurs when multiple pieces of information are known about people, which may be distorted, partially occluded, or disguised, and when the imaging conditions are totally unorthodox! Face Recognition in the Wild has emerged as a field of research in the past few years. Its main purpose is to challenge constrained approaches of automatic face recognition, emulating some of the virtues of the Human Visual System (HVS), which is very tolerant to age, occlusion, and distortions in the imaging process. HVS also integrates information about individuals and adds contexts together to recognize people within an activity or behavior. Machine vision has a very long road to emulate HVS, but face recognition in the wild, using the computer, is a road to perform face recognition in that path. Our research group has been developing a front-end approach for face recognition in the wild, which builds on the state-of-the art in theory and algorithms of facial biometrics and builds upon our own contribution in pose-invariance, illumination modeling and feature optimization. Our approach hinges on two major building blocks: representation and recognition. Representation: constructs a robust representation for faces (from video, still images and other media) into a gallery that is easy to enroll and search. Recognition: constructs an approach to detect faces (from video, still images and other media), and extract expeditiously an optimum feature vector for discriminatory facial key points (around the eyes, nose and lips), suitable for matching with candidate faces in the gallery.

We shall refer to sample references most pertaining to the work proposed in this manuscript. The team has the following contributions: i) Built a front-end Biometric Optical Surveillance System (BOSS) for unconstrained Face Recognition at a Distance (FRAD) up to 150 meters [1-5]; ii) developed a methodology to detect and track multiple faces [6]; iii) developed a

pose-invariant approach for face recognition at a distance [7,8]; iv) developed an image illumination model for generalized lighting and object characteristics, which corrects for illumination variations at random poses [9-12]; v) developed a heat-kernel approach for face recognition suitable for part-based and holistic face recognition [13]; vi) developed face recognition on low-resolution thermal and video imaging [14,15]; vii) developed facial biometric systems for study of autism, involving man-machine interfaces with humanoid robots, non-intrusive vital sign, and expression measurements for behavioral studies of people with special needs [16,17].

The contributions of this paper are:

a. introducing novel representation for face recognition in the wild for video and still imaging addressing the most general view of A-PIE imaging conditions;

b. describing how the representation enables model-based design and evaluation of the three major components of face recognition: face detection, facial feature extraction, and face recognition; and

c. Provide sample evaluations for the proposed representation in an outdoor and indoor settings, using high and low resolution cameras, for Face Recognition at a Distance (FRAD).

Proposed Representation

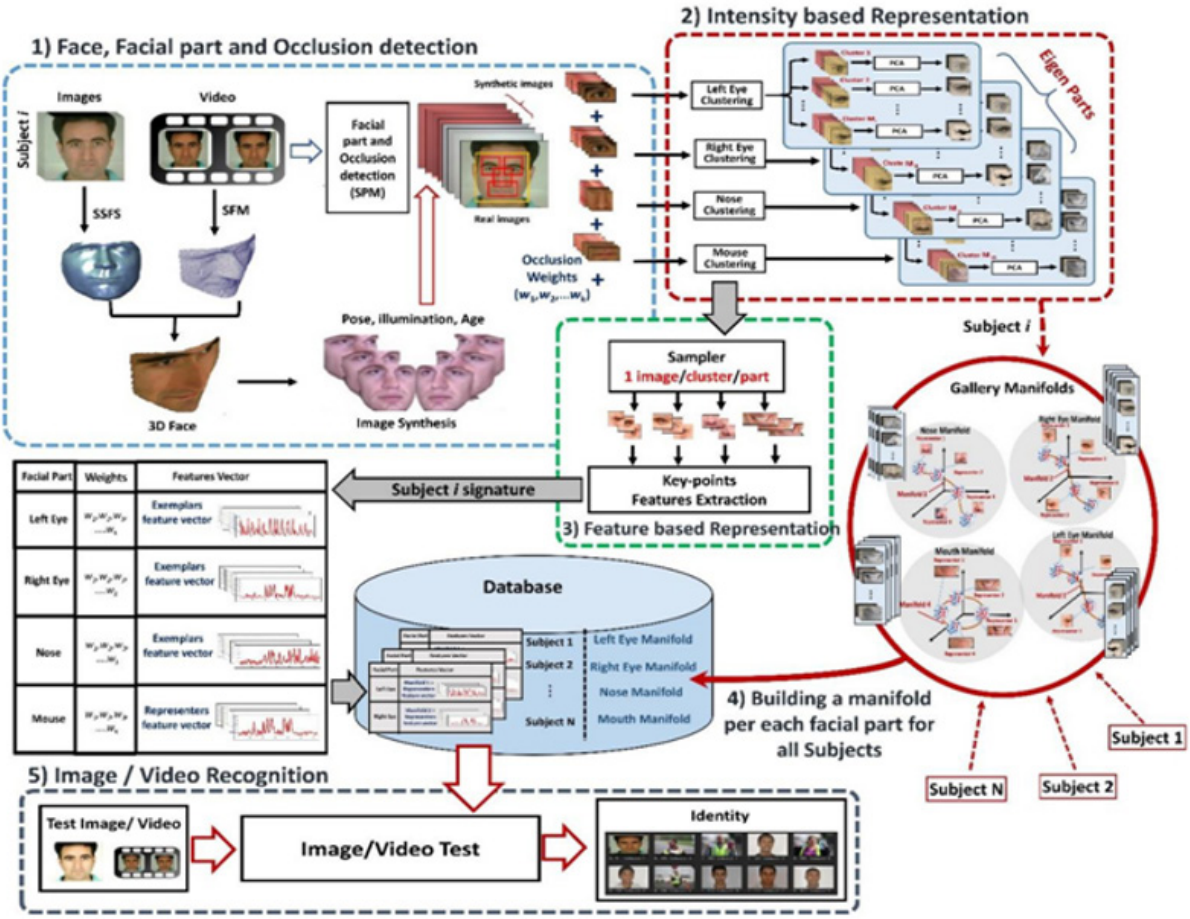

Figure 1: Proposed representation pipeline. From detected facial regions in images or videos, four facial parts (right and left eyes, nose and lip) are extracted from original and novel poses; and their occlusion is given a weight. An elaborate process is used to optimize the features, per part, and assign them a node in the gallery manifold. In the recognition phase, features from detected facial images or videos are detected on the fly, and their features are extracted and used in matching with candidate objects in the gallery manifold using optimal search methods. The framework has five main components, the input/output of every component will be as following: 1) the first component takes the detected faces as an input and generates face parts plus feature points in every part plus occlusion weight. 2) The second component takes the face parts and cluster each part image to fixed number of clusters and generate the cluster's PCA projection for the four face parts. 3) The third component takes the clusters then samples each cluster and extract 2D signatures for each sample with the occlusion weight. 4) The fourth component take the PCA projects and draw the inference on cluster of each face part manifold. 5) The fifth component takes test (probe) image or video and applies the procedure described in the following subsection and search the gallery for the matched identity.

Figure 1 illustrates the proposed representation for face recognition in the wild. Starting from the given data set all the faces will be detected and tracked. We assume that this data will not be complete to describe variation in age, pose, illumination, and expression of the subject under modeling. To complete this data set, we generate an accurate 3D reconstruction, which enables generating novel poses, correct for illumination variations. Two approaches have shown promise for generating the sparse 3D reconstruction: (i) Statistical Shape from Shading (SSFS) which uses databases of shapes and albedos and deploys spherical harmonics and partial least square optimization to generate a 3D reconstruction from a single image [5,9,11]; (ii) Structure from Motion (SFM) (e.g., [18,19]). This route uses all advances in appearance modeling and object reconstruction from a sequence of images in the computer vision literature in the past two decades. Illumination models from Computer vision (e.g., [20,21]), computer graphics and computational photometry (e.g., [22]) will be used for proper modeling of illumination. From detected facial regions extracted from original images or videos and novel generate synthetic images (real+synthetic), four facial parts (right and left eyes, nose and lip) will be extracted using a Selective Part Model (SPM) [6]. The SPM generates ensemble of the left eyes, right eyes, nose and lips. These ensembles are descriptive of various forms of the parts; e.g., close eyes, occluded eyes; occluded nose, smile lips, closed lips, open lips, etc. Hence, per each region, we will have a set of sub-classes { ,∈ [1,4]; = } for each part and their occlusion is given a weight.

We propose to generate an optimal set of images per part sub-classes { ,∈ [1,4]; = } using clustering techniques–this is to minimize the within class scatter and maximize the between class scatter, in each of the part sub-class. The output of this step is an ensemble of parts (the sub-classes per part) with fewer redundancies. They are "images." In this step, we follow a similar approach to See & Swaren [23]. The exemplar technique will select from the original data (if available) over the generated synthetic data (Figure 1). Then we may follow either of two parallel routes to generate the optimal representation:

a. An image-based route will map the inference on clusters of face parts to Grassmann manifold, which captures the global characteristic of the local clusters of face part. We will have four manifolds one for each part.

b. A feature-based route which captures variations within every part clusters, by exemplar sample which can be the mean of each cluster. Then 2D signatures, (e.g., LBP, Gabor wavelet, LOIP, ORB) are reconstructed around each key-point (e.g., Rara [2]) for each sample image.

The features (very large vector) may be optimized by various learning approaches and used to build the gallery together with the four manifolds. Novelty in this representation includes adding the Selective Part Model (SPM) for facial feature extraction [6] and the SFM [19] for sparse facial key point reconstruction. Novelty also includes developing novel learning methods for feature optimization using SVM and Deep Learning, in addition to traditional PCA, and Simulated Annealing.

Recognition

Given an image, video and other media that conveys information about an individual, we extract the facial regions in the image as a whole using home grown approach [24] and extract the facial parts (eyes, nose and lips) using an enhanced selective part model, and a false face reduction mechanism. The SPM [6] and facial features detection play dual roles: jointly enhance the facial detection and reduce false positives and create weighted features for recognition. The facial features will hinge on small patches around nine-key points on which we apply SIFT, SURF, LBP, ORB and other feature descriptors. The SPM model handles natural and self-created occlusions (Figure 2) and creates regions for the right and left eyes, the nose and the lips (Figure 3). Features for each of these parts will be created using various common attributes and geometric characteristics. The features from the descriptors and the SPM will be huge; hence, a learning module is essential to generate optimal set of features, which will be used in matching candidate face representations in the gallery. In all, the entire process is model-based and involves the latest in computer vision, computational photometry, machine learning and database design (e.g., [9,12,13,25]).

Figure 2: Detection of partially occluded faces, by considering sunglasses, caps and hands.

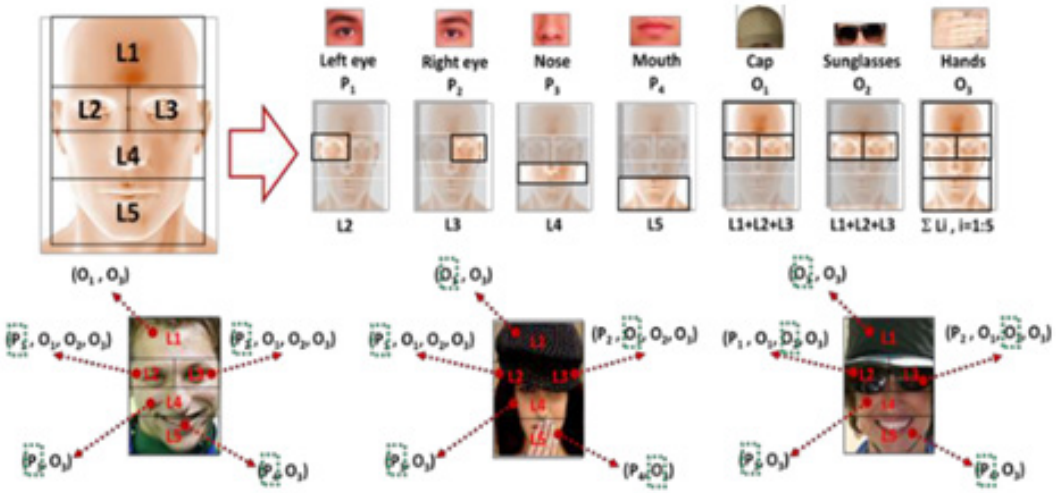

Figure 3: SPM for face detection under partial occlusion by caps, sunglasses and hands.

Evaluation

Gallery reconstruction

Case study: If the given input is a video, all the faces will be detected and tracked in the frames (or assume that the detection and tracking information is given with the provided data). For example, given a five minute video with 30 fps, we will have 5*60*30=9000 frames. More than one subject in the video will be processed in parallel in the following steps (Note: If the subject appears in 2 minutes of a five-minute video, we have 2*60*30=3600 frames.):

a. A 3D model will be generated using the statistical shape from shading (SSFS) [5,11] and the structure from motion [19]. This 3D will enable generating a new set of synthetic images with novel poses, age, and illumination variations.

b. A set of facial feature points (key-points) will be detected in each face for the given and generated images.

c. The detected faces will be divided into four parts or segments of facial regions as: forehead, right and left eyes, nose, and lips (e.g., [11]). The detection step will provide a weight for each face part as a measure for the occlusion level.

d. Cluster the image space of each face part into local clusters using hierarchical agglomerative clustering (STHAC) algorithm [e.g., [23]). If we consider 10 clusters, we will have 360 image in each cluster per face part.

e. The PCA projection of the intensity of part images, and the other is based on the local features of a sample image from each cluster. The 2D signatures, (e.g., LBP, Gabor wavelet, LOIP, ORB) are reconstructed around each key-point (e.g., [2]) for each sample image – That is, we generate 10 feature vectors for each part. These signatures will be enrolled to the database together with the four manifolds. If the given input is a single image or a small set of images, then all the steps above will be the same except the 3D model will be reconstructed using the Statistical Shape from Shading (SSFS) only.

Evaluation on images

a. Staring from detected face, facial feature points are extracted.

b. The facial parts (eyes, nose and lips) are extracted using the Selective Part Models (SPM) approach. The detection step provides a weight as a measure for the occlusion level.

c. 2D signatures, (LBP, Gabor wavelet. LOIP, ORB, etc.), are reconstructed around each key-point [3] for each sample image. Features include LBP, Gabor Wavelet, FAST, BREEF, ORB and LIOP.

d. These features, together with the weights, are used for individual classifiers or in a decision fusion framework to output the final result. Common classifiers include Support Vector Machine (SVM), Dynamic Image to Class Warping (DICW), and minimum distance (e.g., kNN) classifiers.

e. Recognition is obtained by decision fusion from each face part.

Evaluation on video

a. Starting from the face detection and tracking in each frame. A set of facial feature points (key-points) will be detected in each face for the given images (e.g., [24]).

b. The detection step provides a weight as a measure for the occlusion level. A false face reduction mechanism is deployed to reject the false positive faces.

c. Each face part in all the frames is clustered to fixed number of clusters (m-cluster). After clustering we have two routes: one based on the PCA projection of the intensity of part images, and the other is based on the local features of a sample image from each cluster.

d. First route draws inference on cluster of each face part manifold to capture the global characteristic of the local clusters of this part. The first n component (Eigen vectors) of the PCA is

mapped to one point on the part manifold. We have 10 points on each manifold (one point for each cluster). A distance measure to all other enrolled subjects on the manifolds is measured based on the geodesic distance. Then an identity decision is decided from the four manifolds.

e. For the second route, the variations within the local clusters are captured by exemplar sample, which can be the mean of each cluster. Assuming 10 clusters, each cluster has 360 images, selecting one image from each cluster so we will have 10 sample images for each part.

f. 2D signatures, (LBP, Gabor wavelet. LOIP, ORB, etc.), are reconstructed around each key-point (e.g., [2]) for each sample image (10 feature vectors for each part).

g. A sample from each cluster is selected and used for recognition against the gallery as the case of a single image. These provide an identity label from each sample.

h. Recognition is decided using decision fusion techniques.

Results

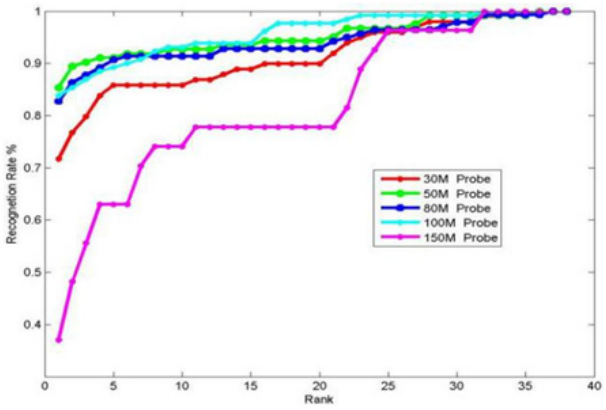

Figure 4: Performance of FRAD setup with Canon EOS 7D: 18MP camera.

The proposed representation in (Figure 1) is modular, enabling various scenarios of face recognition in the wild. Space limitations dictated removal of various performance curves. We only show sample results. Figure 4 shows recognition for the representation in an image-based outdoor face recognition at a distance (The BOSS system [1-5]). The system used Canon EOS 7D: 18MP camera. 11 subjects for five distances: 9 poses (3 for Yaw, Pitch and Roll), outdoor sunny illumination, and modest expression per setting. Gabor and LBP features were used and a kNN distance classifier. Various combinations of facial occlusions were deployed (caps, sunglasses, eyeglasses, makeup, hoodies, etc.). Computation were on 8 CPU machine, un-optimized code; decision took 40sec from enrollment to recognition (Figure 5). Components-wise evaluation of the proposed representation, for pose-invariant face recognition is in [8], and heat-kernel feature matching is in [13]. We also refer to the Master Thesis of the first author [15]. The results demonstrate flexibility of the proposed representation. Current efforts are focused on video FRAD in the wild.

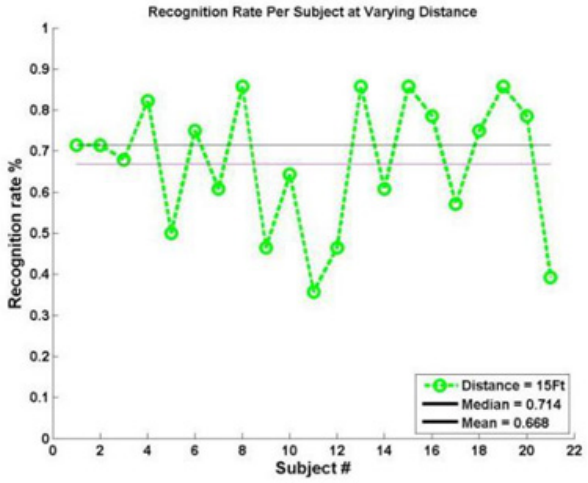

Figure 5: Performance of FRAD setup with iPhone 4: 5 MP camera.

Conclusion

This paper presented a generalized representation for face recognition in the wild showing the major modular building blocks of face detection, signature generation and matching. The representation is the viewpoint of several years of research in this area by our group and is optimal, in many respects, for extreme imaging conditions. It is applicable for still images and video. Software and databases are available and may be requested from the contact author.

References

1. Rara H, Elhabian S, Ali A, Miller M, Starr T, et al. (2009) Face recognition at a distance based on sparse-stereo reconstruction. IEEE CVPR Biometrics Workshop, USA.
2. Rara H, Farag A, Davis T (2011) Model-based 3D shape recovery from single images of unknown pose and illumination using a small number of feature points. International Joint Conference on Biometrics (IJCB), pp. 1-7.
3. Rara H, Elhabian S, Ali A, Miller M, Starr T, et al. (2010) Face recognition at-a-distance using texture and sparse-stereo reconstruction. Proc of IEEE Fourth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1221-1224.
4. Parris J, Wilber M, Helfin B, Rara H, El barkouky A, et al. (2011) Face and eye detection on hard datasets. International Joint Conference on Biometrics (IJCB), pp. 1-10.
5. Rara H, Elhabian S, Starr T, Farag A (2010) 3D face recovery from intensities of general and unknown lightning using partial least squares. Proc of 2010 IEEE International Conference on Image Processing (ICIP), pp. 4041-4044.
6. Elbarkouky A, Farag A (2013) Selective Part Model (SPM) for face detection. CVIP Lab TR-11-2013.
7. Mostafa E, Farag A (2012) Dynamic weighting of facial features for automatic pose-invariant face recognition. Proceedings of Ninth Conference on Computer and Robot Vision, pp. 411-416.
8. Mostafa E, Ali A, Alajlan N, Farag A (2012) An automatic pose invariant face recognition at distance approach. ECCV, Germany.
9. Elhabian S, Rara H, Farag A (2011) Towards accurate and efficient representation of image irradiance of convex-lambertian objects under

unknown near lighting. International Conference of Computer Vision (ICCV), pp. 1732-1737.

10. Elhabian S, Rara H, Farag A (2011) Towards efficient and compact phenomenological representation of arbitrary bidirectional surface reflectance. In Proceedings of the British Machine Vision Conference (BMVC), pp. 89.1-89.11.

11. Elhabian S, Mostafa E, Rara H, Farag A (2012) Non-lambertian model-based facial shape recovery from single image under unknown general illumination. Proceedings of Ninth Conference on Computer and Robot Vision (CRV'12), pp. 252-259.

12. Elhabian S, Farag A (2013) Analytic bilinear appearance subspace construction for modeling image irradiance under natural illumination and non- lambertian reflectance. Proceedings of Computer Vision and Pattern Recognition, CVPR.

13. Abdelrahman M, Farag A, Elmelegy M (2013) Heat front propagation contours for 3D face recognition. Proc of IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS'13), USA.

14. Mostafa E, Hammoud R, Ali A, Farag A (2013) Face recognition based on local descriptors of facial features in low resolution thermal images. Journal of Computer Vision and Image Understanding.

15. Farag M (2013) Face recognition in the wild. Master of Engineering Thesis, University of Louisville, USA.

16. Niese R, Al Hamadi A, Farag A, Neumann H, Michaelis B (2011) Facial expression recognition based on geometric and optical flow features in colour image sequences. British Computer Vision Journal.

17. El Barkouky A, Mahmoud A, Graham J, Farag A (2013) An interactive educational drawing system using a humanoid robot and light polarization. International Conference of Image Processing.

18. Forsyth DA, Ponce J (2002) Computer vision: A modern approach. Prentice Hall Professional Technical Reference, USA.

19. Garg R, Roussos A, Agapito L (2013) Dense variational reconstruction of non-rigid surfaces from monocular video. CVPR, IEEE, USA.

20. Belhumeur PN, Kriegman DJ (1998) What is the set of images of an object under all possible illumination conditions? International Journal of Computer Vision 28: 245-260.

21. Basri R, Jacobs D (2003) Lambertian reflectance and linear subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(2): 218-233.

22. Ramamoorthi R, Hanrahan P (2002) Frequency space environment map rendering. ACM Transactions on Graphics 21(3): 517-526.

23. See J, Eswaran C (2011) Video-based face recognition using spatio-temporal representations. In Tech Open Access Publisher, UK, pp. 1-20.

24. Mostafa E, Farag A (2012) Complex bingham distribution for facial feature detection. Proceedings of European Conference on Computer Vision Workshops (ECCV'12 Workshops), pp. 330-339.

25. Rara H (2011) 3D facial shape estimation from a single image under arbitrary pose and illumination. CVIP Lab, University of Louisville, USA.