2015 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2014, BOSTON, USA

A HIERARCHICAL BAYESIAN NETWORK FOR FACE RECOGNITION USING 2D AND 3D FACIAL DATA

Iman Abbasnejad [1,2]

[1] Queensland University of Technology
2 George Street, Brisbane QLD 4000, Australia

Damien Teney [2]

[2] The Robotics Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, USA

ABSTRACT

In this paper, we tackle the problem of face classification and verification. We present a novel face representation method based on a Bayesian network. The model captures dependencies between 2D salient facial regions and the full 3D geometrical model of the face, which makes it robust to pose variations, and useable in unconstrained environments. We present experiments on the challenging databases FERET and LFW, which show a significant advantage over state-of-the-art methods.

1. INTRODUCTION

The problem of face identification and recognition has been studied for almost 60 years but is still considered unsolved [1]. Face classification is generally approached in two steps. First, due to the high dimensionality of face images, dimensionality reduction algorithms are usually applied to effectively extract facial image features relevant to face classification. Second, a classifier is trained for face recognition using these features. Common choices for the classifier are, for example, Linear Discriminant Analysis (LDA), Support Vector Machines (SVM) or Nearest Neighbor (NN). Humans are exceptionally effective at identifying different human subjects from facial images, but artificial systems still fail to achieve comparable recognition accuracy. It is well understood that the performance of computational systems is affected by a number of factors, arising from image nuisances such as illumination and pixel noise, facial disguise or, importantly, variations of 3D pose. In the traditional approach outlined above, accounting for these nuisances is meant to be achieved by extracting features in the *image space* that would be insensitive to these variations. This is obviously a major challenge that limits the applicability of this paradigm.

Advanced approaches to face recognition can be divided into 2D, 3D, and hybrid models. Although some models using only 2D features perform reasonably well (e.g. [2, 3]), the lacking degree of freedom obviously limits their performance in conditions of large variations of pose, expression, and lighting [4]. To handle these problems, some methods [5] use facial regions, such as the nose and its surrounding area, which are minimally affected by deformations caused by facial expressions. This idea was developed e.g. in [6] using a set of 38 regions that densely cover the face, from which a subset is then selected for classification. Alternatively, 3D models are more robust against illumination changes, but they must cope with additional challenges such as face alignment, missing data and partial occlusions. These models are also more sensitive to facial expressions than their 2D counterparts [7]. The hybrid, or multi-modal models have shown so far the most promising results. By combining the 2D and 3D processing paths in a single architecture, they address the weaknesses of the individual approaches.

In this paper, we introduce a novel method that leverages both 2D and 3D facial features. We use a Bayesian Network (BN) to capture dependencies between 2D salient facial features (namely, the eyebrows, the eyes, the nose, and the mouth) and the full 3D model of the face. As a consequence, the proposed method is robust against changes in illumination and facial pose. The contributions of this paper are twofold.

- We propose a novel strategy for face recognition that uses a Bayesian network to model dependencies among 2D facial regions and the 3D geometrical model of the face.
- We present an evaluation of the method on two challenging facial databases, FERET and LFW, and demonstrate better detection performance than the state-of-the-art.

The paper is organized as follows. After a brief review of related work in Section 2, Section 3 describes the proposed method. Section 4 presents a method for tracking facial landmarks and actually extracting features from the images. In Section 5, we present experiments and comparisons with existing methods.

978-1-4673-7454-5/15/$31.00

2. RELATED WORK

Existing face recognition methods can be categorized into three main different categories: 2D, 3D, and hybrid 2D/3D. The 2D techniques, such as [2, 3, 8, 9, 10], try to classify the face images using purely image-based models. The performance of such methods is very sensitive to variations in expression, to illumination conditions, and to head orientation. They are also typically very sensitive to facial disguise (e.g. a person wearing glasses or a scarf) and to simply misaligned images.

The limitations of 2D face classification methods have supported the belief that effective recognition of identity should be obtained through multi-biometric technologies. In particular, interest has grown on using the geometry of the anatomical structure of the face, rather than its appearance. A number of systems for 3D face matching have thus been developed in the recent years. See for example [11] for a survey.

The existing systems based on 3D models can be further categorized into two distinct subsets: holistic [11, 12, 13] and region-based [14] techniques. The holistic approach employ information from the entire face, or from large regions thereof. Those methods are quite sensitive to proper alignment and to varying facial expression. The region-based techniques mitigate these issues by filtering out regions most affected by facial expression and other spurious causes. Since those methods are based on parts of the faces, their performance depends on the quality of the local features extracted in the corresponding regions.

Hybrid approaches have shown the highest accuracy among face recognition systems, by combining advantages of the 2D and 3D processing paths, although at the cost of a greater complexity. Jahanbin *et al.* [15] used 2D and 3D Gabor coefficients in their face recognition system. In [16], the authors applied PCA to depth and 2D face images separately, then fusing them for face recognition. In [13], ICP registration of a 3D face model was combined with Linear Discriminant Analysis on face images, thus improving over 2D matching alone, in particular in the case of large variations of pose and illumination. Although promising results have been reported with these methods, several additional considerations must be made. One limitation of the previous studies is that they generally did not model or used the dependencies between facial regions. Important facial details are also often missed, e.g. near the mouth, chin and cheeks. For example, considering a face with a smiling expression, the image region containing mouth is certainly related (geometrically, as well as appearance-wise) to those containing the chin and the cheeks. Modeling these dependencies is thus of great interest as such details can prove to be key distinguishing elements. In this work, we address this very issue using a generative model represented as a Bayesian Network, which explicitely captures dependencies between facial parts.

3. PROPOSED MODEL

In this section, we first review the basics of Bayesian networks, and then present its application as our 2D/3D face model.

3.1. Bayesian Networks

A Bayesian Network (BN) is a type of probabilistic graphical model. It is defined as directed acyclic graph, in which the nodes represent random variables $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_N\}$ and the edges represent the dependencies among these random variables. Given the value of its parents, each variable is conditionally independent of its non-descendants. A BN can effectively represent and factor joint probability distributions and is suitable for classification tasks. More specifically, given a set of random variables, the full joint distribution is given by:

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) &= p(\mathbf{x}_1) \times p(\mathbf{x}_2|\mathbf{x}_1) \times \ldots \\ &\quad \times p(\mathbf{x}_N|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N-1}) \\ &= \prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}) \\ &= \prod_{i=1}^{N} p(\mathbf{x}_i|pa(\mathbf{x}_i)) \end{aligned} \quad (1)$$

where $pa(\mathbf{x}_i)$ is the parent of the variable $\mathbf{x}_i$. Many algorithms have been proposed in the literature for learning Bayesian Networks [17, 18]. Some of these may seem attractive, but high computational complexity often limits their use in the case of high dimensional data such as images. For our purposes, we use the method presented in [19] to learn the BN. It selects a structure by minimizing a sequence of two cost functions. The first optimization is over the local error in the log-likelihood domain. This function considers every pair from the set of variables $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, which are independent of each other. In particular, the function organizes the variables into a large number of candidate subsets such that the error is minimized. This step restricts the network to represent dependencies that only occur within subsets. Finally, the model selects a network by minimizing the global measure of empirical classification error computed over the set of input images. Please refer to [19] for additional details.

3.2. Definition of 3D Model and Feature Points

The 3D face model considered in this paper is similar to the 3D Morphable Model (3DMM) [20]. The 3DMM represents a face as a shape vector S_{mod} and a texture vector T_{mod}, which are linear combinations of, respectively, shapes and

textures of the m face examples:

$$\mathbf{S}_{mod} = \sum_{i=1}^{m} \alpha_i \mathbf{S}_i \ , \qquad \mathbf{T}_{mod} = \sum_{i=1}^{m} \beta_i \mathbf{T}_i \tag{2}$$

$$\sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{m} \beta_i = 1 \tag{3}$$

where the shape vectors $\mathbf{S} = (X_1, Y_2, Z_1, \ldots, X_n, Y_n, Z_n)^T \in \mathcal{R}^{3n}$, with X, Y, Z being the coordinates of n vertices, and the texture vectors $\mathbf{T} = (R_1, G_1, B_1, \ldots, R_n, G_n, B_n)^T \in \mathcal{R}^{3n}$ with R, G, B being the color values of n vertices.

In this work, the possible shape deformations are learned with by PCA as follows, using the CASIA-3D FaceV1 Database[1]. Note that the original face images have varying poses, and that the cloud points have missing data. We therefore first fit these faces by a generic 3D model as in [21]. Following this preprocessing, the 3D faces are aligned and have similar parametric forms. We can then apply PCA and get deformable 3D face models composed by the eigenvectors $\mathbf{s}_i$ and $\mathbf{t}_i$ of the covariance matrices. We then represent each shape-vector and texture-vector model as follows:

$$\mathbf{S}_{model} = \bar{\mathbf{S}} + \sum_{i=1}^{m-1} \alpha_i \mathbf{s}_i, \quad \mathbf{T}_{model} = \bar{\mathbf{T}} + \sum_{i=1}^{m-1} \beta_i \mathbf{t}_i, \tag{4}$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ are the mean shape and texture vectors, respectively. The coefficients α and β are drawn from probability distributions estimated from the face examples. More precisely,

$$p(\alpha) \ \propto \ exp(-\frac{1}{2} \sum_{i=1}^{m-1} (\frac{\alpha_i}{\sigma_i})^2) \tag{5}$$

$$p(\beta) \ \propto \ exp(-\frac{1}{2} \sum_{i=1}^{m-1} (\frac{\beta_i}{\delta_i})^2) \tag{6}$$

where σ and δ are the eigenvalues of respectively the shape and texture covariance matrices of the face examples.

3.3. Capturing depencies between 2D and 3D data

The objective of our model is to capture the dependencies between salient facial features and the full 3D geometrical information of the face. However, finding all the dependencies in a model is a NP hard problem. Therefore, we restrict ourselves to a smaller, and thus more efficient model. The model considered in this paper is a Bayesian network with seven visible and ten hidden nodes, as represented in Fig. 1. The set of visible nodes (lower part of Fig. 1) are the 2D salient facial features extracted from the input face images. The set of hidden nodes (upper part of Fig. 1) are the hidden causes that generate the observations. These hidden variables model the

[1]http://biometrics.idealtest.org

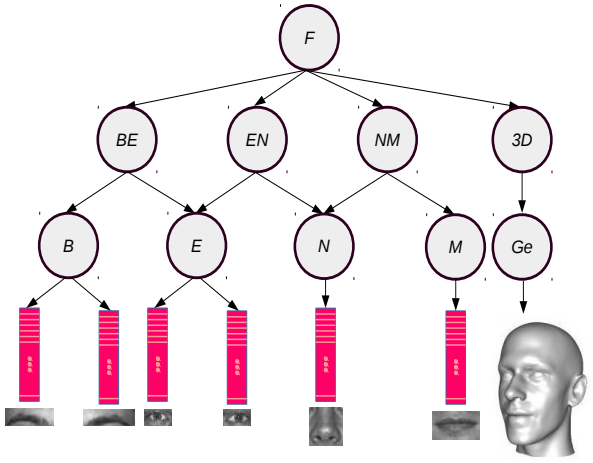


Fig. 1. The Bayesian Network model considered in this paper. B, E, N, M, and Ge refer respectively to eyebrows, eyes, nose, mouth and 3D geometrical information of the face.

relationships between the different parts of the face, namely the eyes, the eyebrows, the nose, and the mouth.

The probability distributions of the hidden nodes are given by discrete values in probability tables. The probability density functions of the distributions of the visible nodes are specified by conditional Gaussians:

$$p(\mathbf{X}_i = \mathbf{x}_i | pa(\mathbf{x}_i = i)) = \mathcal{N}(\mu_{x_i} + \gamma pa(\mathbf{x}_i), \sigma^2_{x_i}(1-\rho^2)) \tag{7}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution of mean μ and standard deviation σ. In Equation 7, the $\mu_{\mathbf{x}_i}$ and $\sigma^2_{\mathbf{x}_i}$ are the mean and variance of the feature x_i, and $pa(\mathbf{x}_i)$ refers to the parents of the node $\mathbf{x}_i$. The ρ is the correlation coefficient between the node x_i and its parents $pa(\mathbf{x}_i)$, and is defined as

$$\rho = \frac{cov(\mathbf{x}_i, pa(\mathbf{x}_i))}{\sigma_{\mathbf{x}_i} \sigma_{pa(\mathbf{x}_i)}} \ .$$

Similarly, γ is defined as

$$\gamma = \frac{cov(\mathbf{x}_i, pa(\mathbf{x}_i))}{\sigma^2_{pa(\mathbf{x}_i)}} \ .$$

4. FACE TRACKING AND FEATURE EXTRACTION

Once the face model is built as described above, appropriate facial features for each visible node need to be extracted from the images. This section describes how this is performed, through face tracking and then feature extraction.

4.1. Facial feature tracking

The facial features are tracked using Constrained Local Models (CLM) [22]. In our case, we use a CLM composed of 66

landmarks distributed along the top of the eyebrows, the inner and outer lip outlines, the outline of the eyes, the jaw, and along the nose.

4.2. Feature extraction

After tracking the individual facial components, a similarity transformation algorithm is applied to the facial features with respect to the normal facial shape. This step normalizes against scale, rotation and translation. This normalization provides further robustness to the effects of head motion. Once the texture is warped into this fixed reference, SIFT descriptors are computed around the outer outline of the mouth, the eyes, the nose, and the eyebrows. Due to the large number of resulting features (128 times the number of points), the dimensionality of the resulting feature vector is reduced using PCA to keep 95% of the variance.

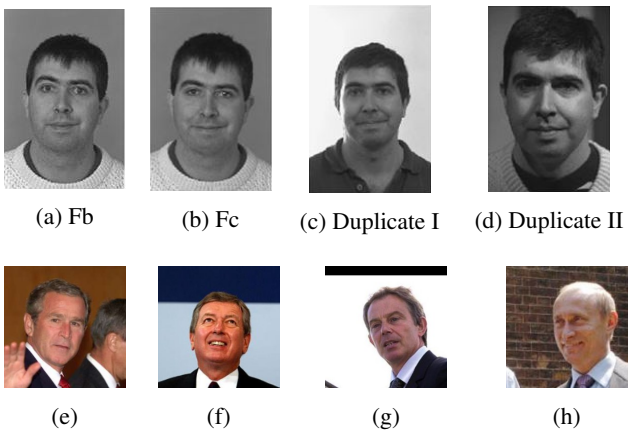
5. EXPERIMENTS

This section describes our experiments on the two face databases FERET and LFW.

5.1. Databases

FERET: This database was used to evaluate the robustness of our system. It has been widely use for evaluating methods for face recognition [23]. It contains images of 1196 individuals with up to 5 images of each individual, captured under different lighting conditions, with varying, non-neutral expressions, and obtained over a period of three years. The complete dataset is partitioned into two disjoint sets: *gallery* and *probe*. The *gallery* set is provided with labels and is used only for training. the *probe* set is used for testing. The *probe* set is further divided into four categories: (I) *Fb Images*, which are similar to the images found in gallery with small variations in expressions; (II) *Fc Images*, which are recorded with different cameras under different lighting conditions; (III) *Duplicate-I Images*, which are taken within the period of 34 months, and finally (IV) *Duplicate-II Images*, which are taken one and a half year later. Fig. 2 shows examples of these different categories.

Labeled Faces in the Wild (LFW): This database was used to further evaluate the robustness of our system when dealing with more variations of imaging conditions. The database [24] contains 13233 face images of 5749 different persons of mixed gender and ages. It is considered to be very challenging as it features face images in various poses, lighting conditions, dressing, etc. The dataset is divided into two subsets named *View 1* and *View 2*. The former is used for training and validation parameters, while the latter is used for testing. Examples of the database are shown in Fig. 2. Note that we used the aligned version of this database as presented in [25].



(a) Fb (b) Fc (c) Duplicate I (d) Duplicate II

(e) (f) (g) (h)

Fig. 2. Example images from the FERET (a–d) and LFW (e–h) databases. Captions (a–d) refer to partitions of the test set, as in columns of Tables 2 and 1 (see text for details).

Method	Area under ROC curve				
	Fb	*Fc*	*Dupl. I*	*Dupl. II*	*LFW*
Proposed	93.15	87.90	70.57	61.27	87.75
Yi *et al.*2013 [26]	83.36	75.47	55.51	60.94	76.92
Heusch *et al.*2010 [27]	78.62	69.50	51.17	52.67	75.90

Table 1. Area under ROC curve on the FERET and LFW databases.

Method	F_1-score				
	Fb	*Fc*	*Dupl. I*	*Dupl. II*	*LFW*
Proposed	68.83	63.81	57.21	52.70	66.70
Yi *et al.*2013 [26]	61.86	59.19	55.87	53.79	65.44
Heusch *et al.*2010 [27]	58.55	55.67	53.18	48.76	62.52

Table 2. F_1-score on the FERET and LFW databases.

5.2. Evaluation settings

To evaluate the performance of our method, we report two measures: the area under the ROC curve and the maximum F_1-score. The F_1-score is defined as: $F_1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$, and conveys the balance between precision and recall. F_1-score is a better performance measure than the area under the ROC curve, as the ROC curve is better suited to evaluate binary classification rather than detection. It does not reflect the effect of the proportion of the positive to negative samples. We evaluated the proposed method with a Matlab and Python implementation.

5.3. Results

Our results are reported in Tables 1 and 2. We compare them with two competing approaches, from Heusch and Mar-

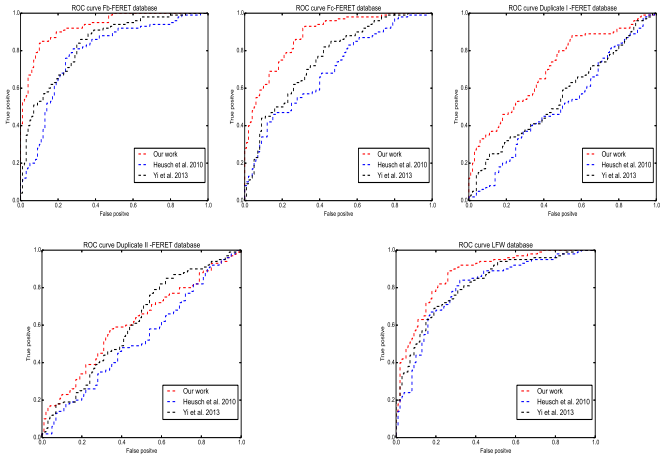


Fig. 3. ROC curve of FERET and LFW databases.

cel [27], and Yi *et al.* [26]. The former, the most similar to our approach, used a BN model with seven hidden and six visible nodes that captured dependencies between 2D facial features extracted from salient facial regions. The authors [27] compared their work with HMM, GMM, Eigenfaces, and Fisherfaces [28] and reported state-of-the-art performance. Despite similitudes, note that our model differs from [27] in that we use a different BN structure. Moreover, we also combine 2D and 3D data to improve robustness, whereas [27] only considered 2D features extracted from specific facial regions.

In [26], the authors tackled the problem of 3D face recognition with a system robust to facial pose and head orientation. As shown in our evaluation (Tables 1 and 2), Heusch *et al.* performs relatively poorly. A possible reason is that classification is only based on features extracted from some facial regions, without consideration to others facial properties, such as shape and geometrical information. Our method consistently outperforms the two compared approaches [27, 26]. Fig. 3 reports the ROC curve on both the FERET and LFW databases.

6. CONCLUSIONS

In this paper, we presented a novel approach to face recognition designed to be robust against changes of pose and illumination, and useable in unconstrained environments. We used a Bayesian network to capture dependencies between 2D salient facial regions and the 3D geometrical model of the face. Experiments on the two challenging databases FERET and LFW showed that the proposed method significantly outperforms the state-of-the-art.

7. REFERENCES

[1] Richard Willing, "Airport anti-terror systems flub tests face-recognition technology fails to flag suspects," *USA TODAY, September 4, 2003. Available at http://www.usatoday.com/usatonline/20030902/5460651s.htm.*

[2] Iman Abbasnejad, M Javad Zomorodian, and Ehsan Tabatabaei Yazdi, "Combination of multi-class svm and multi-class nda for face recognition," in *International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, 2012, pp. 408–413.

[3] Iman Abbasnejad, M Javad Zomorodian, M Amin Abbasnejad, and Hossein Ajdari, "Pose recognition using mixture of exponential family," in *CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)*, 2012, pp. 287–292.

[4] Andrea F Abate, Michele Nappi, Daniel Riccio, and Gabriele Sabatino, "2D and 3D face recognition: A survey," *Pattern Recognition Letters*, vol. 28, no. 14, pp. 1885–1906, 2007.

[5] Kyong I Chang, W Bowyer, and Patrick J Flynn, "Multiple nose region matching for 3D face recognition under varying facial expression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1695–1700, 2006.

[6] Timothy C Faltemier, Kevin W Bowyer, and Patrick J Flynn, "A region ensemble for 3-d face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 62–73, 2008.

[7] NM Brooke and Quentin Summerfield, "Analysis, synthesis, and perception of visible articulatory movements.," *Journal of phonetics*, 1983.

[8] Gregory Shakhnarovich and Baback Moghaddam, "Face recognition in subspaces," in *Handbook of Face Recognition*, pp. 19–49. Springer, 2011.

[9] Conrad Sanderson, Ting Shang, and Brian C Lovell, "Towards pose-invariant 2D face classification for surveillance," in *Analysis and Modeling of Faces and Gestures*, pp. 276–289. Springer, 2007.

[10] Matthew A Turk and Alex P Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition (CVPR)*, 1991, pp. 586–591.

[11] Kevin W Bowyer, Kyong Chang, and Patrick Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition," *Computer vision and image understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[12] Gang Pan, Shi Han, Zhaohui Wu, and Yueming Wang, "3D face recognition using mapped depth images," in *Workshops at Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 175–175.

[13] Xiaoguang Lu and Anil K Jain, "Deformation modeling for robust 3D face matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1346–1357, 2008.

[14] Jamie A Cook, Vinod Chandran, and Clinton B Fookes, "3D face recognition using log-gabor templates," 2006.

[15] Sina Jahanbin, Hyohoon Choi, and Alan C Bovik, "Passive multimodal 2D+3D face recognition using gabor features and landmark distances," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 4, pp. 1287–1304, 2011.

[16] Kyong I Chang, Kevin W Bowyer, and Patrick J Flynn, "An evaluation of multimodal 2D+3D face biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 619–624, 2005.

[17] Gregory F Cooper and Edward Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine learning*, vol. 9, no. 4, pp. 309–347, 1992.

[18] Nir Friedman and Daphne Koller, "Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks," *Machine learning*, vol. 50, no. 1-2, pp. 95–125, 2003.

[19] Henry Schneiderman, "Learning a restricted bayesian network for object detection," in *Computer Vision and Pattern Recognition (CVPR)*, 2004, vol. 2, pp. II–639.

[20] Volker Blanz and Thomas Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[21] John D Bustard and Mark S Nixon, "3D morphable model construction for robust ear and face recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2582–2589.

[22] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic, "Robust discriminative response map fitting with constrained local models," in *Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3444–3451.

[23] P Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[24] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[25] Lior Wolf, Tal Hassner, and Yaniv Taigman, "Similarity scores based on background samples," in *Computer Vision–ACCV 2009*, pp. 88–97. Springer, 2010.

[26] Dong Yi, Zhen Lei, and Stan Z Li, "Towards pose robust face recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3539–3545.

[27] Guillaume Heusch and Sebastien Marcel, "A novel statistical generative model dedicated to face recognition," *Image and Vision Computing*, vol. 28, no. 1, pp. 101–110, 2010.

[28] Simon Lucey and Tsuhan Chen, "A gmm parts based face representation for improved verification through relevance adaptation," in *Computer Vision and Pattern Recognition (CVPR)*, 2004, vol. 2, pp. II–855.