Facial Shape Localization Using Probability Gradient Hints

Zhiheng Niu, Shiguang Shan, *Member, IEEE*, and Xilin Chen, *Member, IEEE*

***Abstract*—This letter proposes a novel method to localize facial shape represented by a series of facial landmarks. In our method, the problem of facial shape localization is formulated with a Bayesian inference. Specifically, given a face image, the posterior probability of the facial shape is naturally decomposed into two parts: the likelihood function of local textures and the prior constraints of global shape. The former is provided by the landmark detectors, while the latter is evaluated based on the global shape statistics. The global shape is iteratively estimated in the Maximum A Posteriori (MAP) procedure which is derived in a Lucas-Kanade manner over the probability distribution. Intuitively, in each step, the landmarks are driven by the probability gradient and converge towards the positions which maximize the posterior probability. Experiments on two public databases (XM2VTS and BioID) show the effectiveness of the proposed method.**

***Index Terms*—Boosting, facial shape localization, maximum a posteriori estimation, probability gradient hints.**

I. Introduction

In many computer vision and image understanding tasks, the localization and alignment of a target object within an image is of great importance. Especially in face perception related research areas, facial shape localization, which provides the correspondence of facial landmarks between different face images, has received significant attention because it is one of the key steps in face recognition, face tracking, pose estimation and so on. To solve this problem, many methods have been proposed in recent years, e.g., active contour models (snake) [1], deformable template [2], elastic bunch graph matching [3], Active Shape Model (ASM) [4], and Active Appearance Model (AAM) [5], [6] etc. Among these methods, ASM and AAM, both based on the statistical point distribution model, have been recognized as the most successful ones. To pursue further improvement, a variety of methods have been proposed, generally, in three aspects:

Manuscript received May 07, 2009; revised June 10, 2009. First published June 30, 2009; current version published July 29, 2009. This work was supported in part by the NSFC under Contracts 60832004, 60833013, and 60533030, and also by the National Basic Research Program of China (973 Program) under Contract 2009CB320902. This work was performed at the Institute of Computing Technology, CAS. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dimitri Androutsos.

Z. Niu is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China (e-mail: zhniu@jdl.ac.cn).

S. Shan and X. Chen are with Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, China (e-mail: sgshan@jdl.ac.cn; xlchen@jdl.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2009.2026457

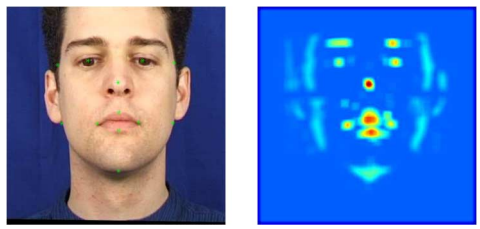

Fig. 1. Some facial landmarks on a face image and their probability distributions predicted by the landmark detectors.

1) More complicated local texture models, consisting of texture representations and feature extractions, are proposed. For example, Gabor features [7] and Haar-like features [8], [9] are combined with the scheme of ASM. In tensor-based AAM [10], tensor representation is adopted to model the large variations of face appearances and generate the AAM basis vectors which are appropriate for the input image.
2) More sophisticated global shape models are adopted. For instance, GMM is deployed in [11]. In [12], the Gaussian Markov Random Field is adopted to model the shape. Besides, part-based shape models are developed in [13].
3) The relationship between local textures and global shape were further formulated and some advanced optimization approaches are utilized. For instance, in the direct appearance models [14], a linear method is directly brought forward to describe the relationship between shape and texture information. [15] and [16] provide robust likelihood evaluations in an MAP procedure. The procedure converges in a principled manner due to the monotonously increased posterior probability in each step. However these methods often need to determine the number of candidate positions of a landmark beforehand. The probability distributions around the candidate positions are often assumed to be Gaussian, which is hardly satisfied in real world applications. For example, as shown in Fig. 1, the contour landmarks are probably located along the image edge and the eye centers are most likely located at the dark areas (e.g., eyebrow) in the image, therefore their probability distributions are not Gaussian.

Motivated by the previous works, in this paper, we propose to solve the problem of facial shape localization under the probabilistic framework. Specifically, an optimization method is exploited to maximize the Bayesian posterior probability of facial shape which consists of two parts: one is the likelihood function of local textures and the other is the prior probability of global shape. In order to accurately predict the former, i.e., the probability distributions of facial landmarks on the image, landmark detectors are trained based on a boosting method [17] using Haar-like features [8], [9], [18]. Unlike previous methods, no

candidate positions are assumed in our method and the probability distribution of a landmark is not assumed to be Gaussian. For the second part, the prior shape probability is evaluated according to the shape statistics obtained via Principle Component Analysis (PCA) form a training set. The optimization procedure is directed by a relative probability gradient under the landmark distribution, which is named as Probability Gradient Hint (PGH). The PGH is computed numerically and thus the shape is updated analytically in each iteration. In other words, the landmarks are driven by the PGHs and converge towards the positions which maximize the posterior probability.

The remaining parts of this paper are organized as follows: Section II gives the description of the probability inference and the detailed specification of the probability of global shape and local textures. In Section III, the substantial optimization procedure is interpreted. Extensive experiments are conducted in Section IV, followed by the conclusions in the last section.

II. Bayesian Inference and Specifications

Similar to ASM/AAM, the shape is denoted as a vector $\mathbf{S} = [x_1\, y_1\, x_2\, y_2\, \ldots\, x_n\, y_n]^T$ where (x_l, y_l) is the location of the l^{th} landmark in the target image and n is the number of landmarks. Given an image $\mathbf{I}$, the goal of pursuing the most likely shape $\mathbf{S}^*$ can be formulated as maximizing the posterior probability $P(\mathbf{S}|\mathbf{I}) = P(\mathbf{S})P(\mathbf{I}|\mathbf{S})/P(\mathbf{I})$. $P(\mathbf{I})$ is a constant independent of $\mathbf{S}$. Furthermore, the local textures from different facial landmarks are assumed to be independent of each other, therefore the optimal shape can be pursued as

$$\mathbf{S}^* = \arg\max_{\mathbf{S}\in\mathbb{S}} P(\mathbf{S}) \prod_{l-1}^{n} P(\mathbf{I}_l|\mathbf{S}_l) \tag{1}$$

where $\mathbb{S}$ is the shape space, $\mathbf{S}_l$ is the position (x_l, y_l), and $\mathbf{I}_l$ is the local texture around (x_l, y_l).

$P(\mathbf{S})$ *Specification:* The shape is controlled by two types of parameters when it changes in a subspace. One is the registration parameter, i.e., similarity transformations including rotation, scaling, and translation; the other is the shape parameter which is presented by the PCA coefficients in the tangent shape space. The registration parameter is denoted as $\theta = [a, b, tx, ty]^T$ and the shape parameter is denoted as $\beta = [c_1\, c_2\, \ldots\, c_m]^T$. The transformed shape can be represented as

$$\mathbf{S}' = \overline{\mathbf{S}'} + \mathbf{V}\beta \tag{2}$$

where $\mathbf{V}$ is the projection matrix and $\mathbf{S}' = [x'_1\, y'_1\, x'_2\, y'_2\, \ldots\, x'_n\, y'_n]^T$ is the transformed shape that best fits the mean shape $\overline{\mathbf{S}'}$ in the least square sense. The transformation of each landmark is represented as

$$\begin{pmatrix} x_l \\ y_l \end{pmatrix} = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \begin{pmatrix} x'_l \\ y'_l \end{pmatrix} + \begin{pmatrix} tx \\ ty \end{pmatrix}. \tag{3}$$

Since $\mathbf{S}$ is determined by θ and β which can be safely assumed to be independent, we have $P(\mathbf{S}) = P(\beta)P(\theta)$. Due to the uniform distribution of $P(\theta)$ and the normal distribution of PCA coefficients, $P(\mathbf{S})$ is proportional to $e^{-(1/2)\beta^T\Lambda^{-1}\beta}$, i.e., $P(\mathbf{S}) \propto e^{-(1/2)\beta^T\Lambda^{-1}\beta}$, where Λ is the diagonal matrix that contains the first m leading eigenvalues derived from PCA of the transformed shapes.

$P(\mathbf{I}_l|\mathbf{S}_l)$ *Specification:* We approximate this conditional probability by integrating the output of the corresponding landmark detector. Specifically, for each landmark, we learn a GentleBoost-based detector [17] based on Haar-like features. For each pixel (x, y) in the target image, if its surrounding local texture (a rectangle image patch centered at (x, y)) is classified by the l^{th} detector as positive, we assign the probability of l^{th} landmark at (x, y) as $P_l(x, y) = 1/N_l$, where N_l is the total number of the positive detections for l^{th} detector in the whole image. Otherwise, a very small positive constant ϵ is assigned, i.e., $P_l(x, y) = \epsilon$.

Algorithm 1: Parameter optimization

Input: The coarsely aligned face image $\mathbf{I}$ (e.g., by a face detector), shape statistics Λ, $\mathbf{V}$.

Output: Optimized facial shape, $\mathbf{S}^*$.

1 Start with the registration parameter $\theta \Leftarrow (1\ 0\ 0\ 0)^T$, shape parameter $\beta \Leftarrow \mathbf{0}_{m\times 1}$.

2 Set $k \Leftarrow 0$, $\mathbf{S}^0 \Leftarrow \mathbf{0}_{2n\times 1}$

3 **repeat** // `Optimization of registration parameter` θ

4 $k \Leftarrow k + 1$

5 Update current shape $\mathbf{S}^k$ via (3) and (2).

6 Assign the probability distribution at (x_l, y_l) and its neighbors by each landmark detector.

7 Compute PGHs via (9).

8 Compute $\Delta\theta$ via (6).

9 Update the registration parameter $\theta \Leftarrow \theta + \lambda_1\Delta\theta$.

10 **until** $\|\mathbf{S}^k - \mathbf{S}^{k-1}\| < \epsilon_1$;

11 Set $k \Leftarrow 0$, $\mathbf{S}^0 \Leftarrow \mathbf{0}_{2n\times 1}$

12 **repeat** // `Optimization of shape parameter` β

13 Operate as $4 \sim 7$.

14 Compute $\Delta\beta$ via (8).

15 Update the shape parameter $\beta \Leftarrow \beta + \lambda_2\Delta\beta$.

16 **until** $\|\mathbf{S}^k - \mathbf{S}^{k-1}\| < \epsilon_2$;

17 Return the optimized shape $\mathbf{S}^* \Leftarrow \mathbf{S}^k$.

III. Parameter Optimization

With the above specifications, the objective function defined in (1) can be rewritten as follows (after performing the natural logarithm):

$$F(\beta, \theta) = \sum_{l=1}^{n} \ln\left(P_l(x_l, y_l)\right) - \frac{1}{2}\beta^T\Lambda^{-1}\beta. \tag{4}$$

The optimization of this objective function is summarized in Algorithm 1, where λ_1 and λ_2 are the steps which control the convergence speed. Briefly speaking, Algorithm 1 consists of two procedures, as described as follows.

Optimization of Registration Parameter θ: As θ has no effect on the second term, by considering β as a constant, the aim is turned to maximize the first term of F. A gradient ascent method

is employed to iteratively optimize the registration parameter. The first-order partial derivative of F w.r.t. θ is

$$\frac{\partial F}{\partial \theta} = \sum_{l=1}^{n} \frac{\nabla P_l}{P_l} \frac{\partial (x_l, y_l)}{\partial \theta} \tag{5}$$

where P_l and ∇P_l are the probability of the l^{th} landmark at (x_l, y_l) and its gradient respectively. According to (3), $\partial(x_l, y_l)/\partial\theta = \begin{pmatrix} x'_l & y'_l & 1 & 0 \\ y'_l & -x'_l & 0 & 1 \end{pmatrix}$, therefore the increment of registration parameter can be computed by

$$\Delta\theta = \sum_{l=1}^{n} \begin{pmatrix} x'_l & y'_l & 1 & 0 \\ y'_l & -x'_l & 0 & 1 \end{pmatrix}^T \left(\frac{\nabla P_l}{P_l} \right)^T. \tag{6}$$

Optimization of Shape Parameter β: The second term of F is second-ordered which facilitates the Gauss-Newton solution. By taking a first-order Taylor expansion of the first term of F at β and applying the chain rule, we have

$$F(\beta + \Delta\beta, \theta) = \sum_{l=1}^{n} \ln\left(P_l(x_l, y_l)\right) + \sum_{l=1}^{n} \frac{\nabla P_l}{P_l} \frac{\partial (x_l, y_l)}{\partial (x'_l, y'_l)} \times \frac{\partial (x'_l, y'_l)}{\partial \beta} \Delta\beta - \frac{1}{2}(\beta + \Delta\beta)^T \Lambda^{-1} (\beta + \Delta\beta) \tag{7}$$

where $\partial(x_l, y_l)/\partial(x'_l, y'_l) = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$ and $\partial(x'_l, y'_l)/\partial\beta = \begin{pmatrix} \mathbf{v}_{2l-1} \\ \mathbf{v}_{2l} \end{pmatrix}$ (i.e., the corresponding rows of the projection matrix $\mathbf{V}$ according to (2)). By setting the derivative of $F(\beta + \Delta\beta, \theta)$ w.r.t. $\Delta\beta$ to zero, we obtain

$$\Delta\beta = \Lambda \sum_{l=1}^{n} \left(\mathbf{v}_{2l-1}^T \mathbf{v}_{2l}^T\right) \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \left(\frac{\nabla P_l}{P_l} \right)^T - \beta. \tag{8}$$

Probability Gradient Hints: Both (6) and (8) include the term $(\nabla P_l / P_l)^T$ which is termed as the Probability Gradient Hint (PGH) of the l^{th} landmark and denoted as $\mathbf{H}_l$. It is computed by

$$\mathbf{H}_l = \begin{pmatrix} \frac{P_l(x_l+1, y_l) - P_l(x_l-1, y_l)}{2P_l(x_l, y_l)} \\ \frac{P_l(x_l, y_l+1) - P_l(x_l, y_l-1)}{2P_l(x_l, y_l)} \end{pmatrix} \tag{9}$$

for $l = 1, 2, \ldots, n$. It can be found that the PGH comes from the image evidence and acts as the relative probability gradient at (x_l, y_l) which directs the moving direction for the consequent iteration. Since $\mathbf{H}_l$ is computed relatively (i.e., $\mathbf{H}_l$ does not change if P_l is multiplied by a constant), therefore the probability mass function, defined in Section II, can be simplified as $P_l(x, y) = 1$ regardless of the detection number (i.e., N_l) on the whole image due to the division operation. To ensure the continuity and smoothness of the probability distribution, P_l is further filtered by a Gaussian window. Benefiting from this property, to compute $\mathbf{H}_l$, we do not need to collect image evidences within the whole image but only within a small neighborhood of the current position (x_l, y_l).

As illustrated in Fig. 2, the $\mathbf{H}_l$ points in the direction of the greatest increase of the logarithm likelihood predicted by the l^{th} landmark detector. By applying (6) and (8), the final decision is made for all landmarks. When the algorithm converges, the PGHs are in balanced state considering the global shape constraint.

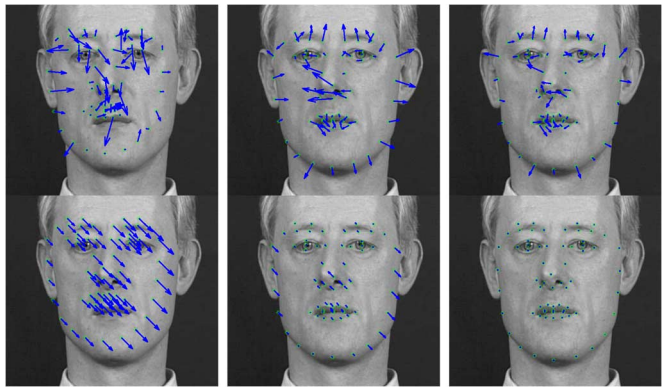

Fig. 2. PGHs (first row) and the predicted moving direction $\Delta\mathbf{S}$ (second row) in three iterations: the beginning of registration parameter θ optimization (left column); the end of θ optimization, i.e., the beginning of β optimization (middle column); the end of β optimization (right column). The magnitudes and orientations of PGHs are represented by the lengths and directions of the blue arrows in the first row (similarly for $\Delta\mathbf{S}$ in the second row).

Discussion: Differences From Lucas-Kanade (LK) Method: As can be seen, our optimization procedure is performed in an LK-like manner [6]. However, there are two differences between the LK method and ours. Firstly, the objective function in our method is the posterior probability function of the shape rather than the squared loss function of image intensity as in the LK method. Secondly, the LK method requires computing the first-order derivative of image intensity in the whole face image, while in our method it is replaced by PGHs computed within the neighbourhoods of the current positions.

IV. Experiments

Experiments are conducted on two public face databases, XM2VTS [19] and BioID [20] to evaluate the proposed method. In XM2VTS, images are acquired under four sessions, and there are 590 face images for each session. The images taken from the first two sessions are used for training, and those in the last two sessions are used for testing. More challenging test is conducted on the BioID database (totally 1521 face images), since no images in the BioID database are used for training and the testing images contain complex pose, lighting and background variations. Each face image in XM2VTS database is released with 68 manually labeled landmarks, and the face image in BioID has 20 manually labeled landmarks.

In the training stage, all the face images are normalized according to the provided eye centers to make the distance between them be 100 pixels. For the boosting algorithm, the positive samples are extracted in a square window of size 25×25 centered at each landmark, and the negative samples are randomly extracted some distance away from the positive ones. The proportion of their amount is 1:5, i.e., each landmark detector is trained from 1,180 positive samples and 5,900 negative samples. For each detector, 100 Haar-like features are selected to build a strong classifier by using the GentleBoost method [17]. In the fitting stage, the input images are automatically localized by a face detector based on the cascaded AdaBoost architecture [18] and coarsely normalized to the same size as those in the training set.

The localization error is evaluated as

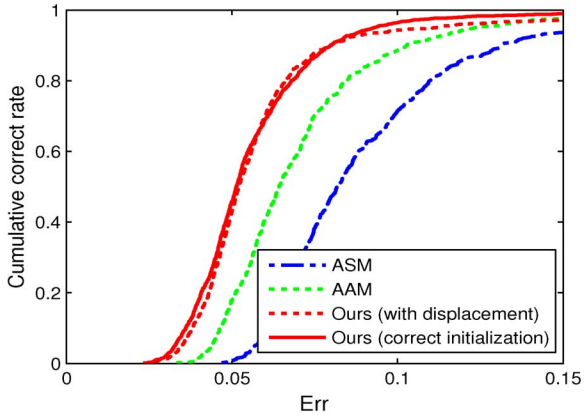

Fig. 3. Comparisons with ASM[1] and AAM[2] on XM2VTS database.

[1]The ASM is implemented by us.

[2]The AAM results are cited from IDIAP homepage [19].

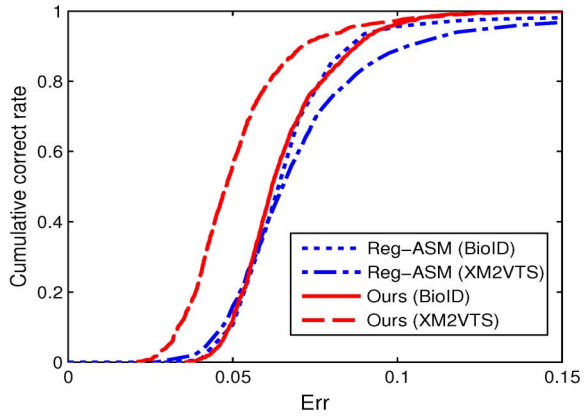

Fig. 4. Comparisons with Reg-ASM[3] on XM2VTS and BioID databases.

[3]The Results of Reg-ASM are cited directly from [9].

$$Err = \frac{1}{n\|\mathbf{S}_L - \mathbf{S}_R\|}\sum_{l=1}^{n}\|\mathbf{S}_l - \mathbf{S}_l^*\| \tag{10}$$

where $\mathbf{S}_L$, $\mathbf{S}_R$ and $\mathbf{S}_l$ are the manually labeled positions of the left eye, the right eye and the l^{th} landmark respectively, $\mathbf{S}_l^*$ is the optimized location for the l^{th} landmark. The performance of each method is plotted as the curve of $z = p(Err < e)$, the proportion of localizations with Err smaller than e against the total number of testing images. Hereinafter, z is called cumulative correct rate.

First, the performance of our method is compared with those of ASM and AAM on the XM2VTS database using the 68 landmarks. The comparison results are shown in Fig. 3. From the figure, it is clear that our method outperforms both ASM and AAM. In the figure, we also report results of our method when the initial shape is purposively made farther away by translating 15 pixels to the left and 15 pixels to the top. Clearly, our method can still converge correctly in such conditions.

Then, as shown in Fig. 4, we also compare our method with the recently proposed method Reg-ASM [9]. It can be seen from the figure that our method outperforms Reg-ASM on XM2VTS database and performs comparably on the BioID database.

V. Conclusions

With Bayesian inference, this letter proposed a new facial shape localization method. The posterior probability is composed of two parts: one is from the image evidence provided by the landmark detectors; the other is obtained from the statistics of the global shape constraints. An MAP procedure is employed to update the shape iteratively by optimizing two types of parameters, i.e. the registration parameters and the shape parameters. Both of them are solved analytically with the PGHs computed numerically. The PGHs lead the landmarks to the most likely places while the global shape constraints ensure the landmarks move in a reasonable manner. Experiments on the XM2VTS and BioID database showed the accuracy and robustness of our approach.

References

[1] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.
[2] A. Yuille, "Deformable templates for face recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 59–70, 1991.
[3] L. Wiskott, J. Fellous, N. Kuiger, and C. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, 1997.
[4] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models—Their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.
[5] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *Proc. Eur. Conf. Computer Vision*, 1998, pp. 484–498.
[6] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, 2004.
[7] F. Jiao, S. Li, H. Shum, and D. Schuurmans, "Face alignment using statistical models and wavelet features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, vol. 1, pp. 321-237.
[8] L. Zhang, H. Ai, S. Xin, C. Huang, S. Tsukiji, and S. Lao, "Robust face alignment based on local texture classifiers," in *Proc. IEEE Int. Conf. Image Processing*, 2005, vol. 2, pp. 354–357.
[9] D. Cristinacce and T. Cootes, "Boosted regression active shape models," in *Proc. Brit. Machine Vision Conf.*, 2007, pp. 880–889.
[10] H. Lee and D. Kim, "Tensor-based active appearance model," *IEEE Signal Process. Lett.*, vol. 15, pp. 565–568, 2008.
[11] T. Cootes and C. Taylor, "A mixture model for representing shape variation," *Image Vis. Comput.*, vol. 17, no. 8, pp. 567–573, 1999.
[12] L. Gu, E. Xing, and T. Kanade, "Learning GMRF structures for spatial priors," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–6.
[13] Y. Huang, Q. Liu, and D. Metaxas, "A component based deformable model for generalized face alignment," in *Proc. IEEE Int. Conf. Computer Vision*, 2007.
[14] X. Hou, S. Li, H. Zhang, and Q. Cheng, "Direct appearance models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. 828–833.
[15] Y. Zhou, L. Gu, and H. Zhang, "Bayesian tangent shape model: Estimating shape and pose parameters via Bayesian inference," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, vol. 1, pp. 109–116.
[16] S. Yan, X. He, Y. Hu, H. Zhang, M. Li, and Q. Cheng, "Bayesian shape localization for face recognition using global and local textures," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 102–113, 2004.
[17] J. Friedman, T. Hastie, and R. Tibshirani, "Special invited paper. Additive logistic regression: A statistical view of boosting," *Ann. Statist.*, vol. 28, no. 2, pp. 337–374, 2000.
[18] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[19] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. Int. Conf. Audio and Video-Based Biometric Person Authentication*, 2003, pp. 72–77.
[20] O. Jesorsky, K. Kirchberg, and R. Frischholz, The Bioid Face Database 2001 [Online]. Available: http://www.bioid.com/downloads/facedb/index.php