Estimating Face Pose by Facial Asymmetry and Geometry

Yuxiao Hu[†], Longbin Chen[*], Yi Zhou[†], Hongjiang Zhang[†]
[†]*{i-yuxhu, i-yizhou, hjzhang}@microsoft.com, Microsoft Research, Asia, Beijing, China,100080*
[*]*lbchen@umsis.miami.edu, University of Miami, Coral Gables, FL, USA, 33124-0620*

Abstract

A robust pose estimation approach is proposed by combining facial appearance asymmetry and 3D geometry in a coarse-to-fine framework. The rough face pose is first estimated by analyzing the asymmetry of the distribution of the facial component detection confidences on an image, which actually implies an intrinsic relation between the face pose and the facial appearance. Then, this rough face pose, as well as error bandwidth, is utilized into a 3D-to-2D geometrical model matching to refine the pose estimation. The proposed approach is able to track a face with fast motion in front of cluttered background and recover its pose robustly and accurately in real-time. Experiment results are provided to demonstrate its efficiency and accuracy.

1. Introduction

Classical 2D face recognition methods are sensitive to pose variation and their accuracies drop when the training and testing faces are not in the same pose so that it is difficult to align them precisely [1][2]. Therefore, extracting face pose parameters, i.e. the three rotation angles α, β, γ in yaw, pitch and roll direction, the 2D position translation (*x, y*) on image and the scale factor *s*, is necessary for further face analysis and recognition. Vetter has proposed a 3D face alignment approach which is able to reconstruct the 3D face model with a single face image under arbitrary pose [3], but the approach is too slow to be used for real-time face tracking and pose estimation. Fast, robust and accurate face pose estimation remains a challenging computer vision problem.

Previous face pose estimation works can be roughly classified into two main categories: *classification-based* approaches and *geometry-based* approaches. *Classification-based* approaches treat face pose as some intrinsic dimensions of the whole face appearance space and attempt to recover the relationship between face pose and its appearance by statistical learning algorithms.[4][5][6][7] However, face appearance space is formed by face samples which coupled many factors such as illumination, expression and human identification besides pose variation, so that above learning methods require large number of training samples of different people under various conditions. But in reality, this kind of data is quite difficult to be collected and labeled. Moreover, accuracies of these methods will be largely limited by the quantity the partition strategy of training data because the majority of these methods utilize interpolation. As results, the face pose estimation error is typically no less than 10º [5][6][7]. On the other hand, *Geometry-based* approaches make use of the 3D structure of human face. The majority of them builds 3D models for human faces and attempt to match the facial features such as the face contour and the facial components of the 3D face model with their projection on the 2D test image. Alter *et. al.* proposed to recover face pose parameters under orthographic projection after the facial feature points are detected [8]. Huang *et. al.* proved the uniqueness of 3D face pose under weak perspective where the face plane is determined by the direction and ratio of the major and minor axes of the ellipse formed by three non-collinear facial key points [9]. These *geometry-based* approaches typically share some common advantages. First, the geometry structure of face reveals the relation between the 3D face pose and the 2D face image which is clearly proved by projection model. Second, they are simple for implementation and the accuracy is prominent when the facial features are located precisely. Third, when more facial features are available, their noises will be reduced by the geometric constrain of the face model [10]. On the other hand, the limitation of the *geometry-based* approaches is that they are sensitive to the location errors of the facial feature points, while facial feature localization remains an open problem. Although [13] and [14] proposed robust and accurate 2D face alignment algorithms, they are only available on frontal pose.

In order to address aforementioned issues, a coarse-to-fine pose estimation framework is proposed in this

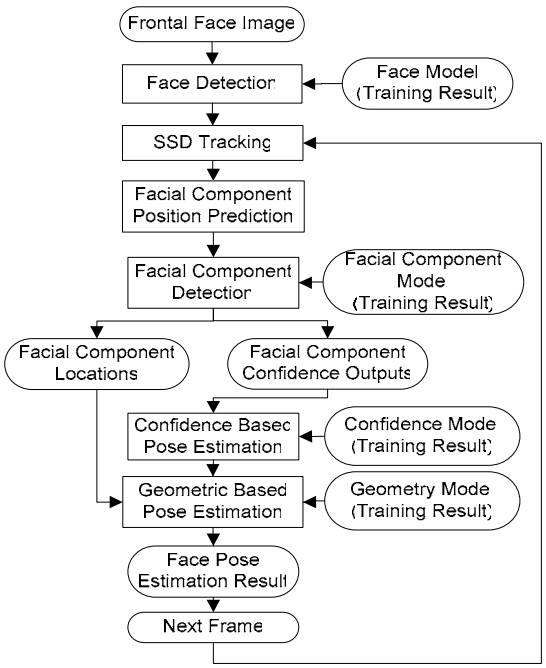

Figure 1. System framework

paper with following highlights: (1) A novel pose estimation method is first used to recover the coarse face pose based on the confidence outputs of facial component detectors. The asymmetry of the confidences distribution actually implies an intrinsic relation between the face pose and the facial appearance. Based on the coarse pose estimation and the facial component detection results, linear regression is conducted to refine and stabilize the pose parameters iteratively. (2) The crucial facial feature localization problem is overcome by combining the geometry face model and the confidence output from facial component detectors. (3) The proposed approach is able to track a face and recover its pose accurately and robustly in real-time in front of cluttered background. The overall accuracy is about 6.54 degree in yaw direction and 5.35 degree in tilt direction, which is acceptable for the requirement of a user attention tracking system.

The framework of the proposed coarse-to-fine face pose estimation is illustrated in Figure 1.

The rest of this paper is organized as follows. Section 2 describes the method of facial component detection with confidence output. Confidence based coarse pose estimator and linear regression based pose estimator are detailed in Section 3 and 4 respectively. Experiment results and system implementation are given in Section 5 to justify the proposed approach. The paper is concluded in Section 6 with some further discussions.

2. Facial component detection

2.1 Initialization

First, in the initialization step, a robust AdaBoost-based face detector is utilized to find a frontal face region [16]. At the same time, five facial components: left and right eye centers, left and right mouth corners and nose are detected as in [15]. Since the face is in frontal pose, the regions containing the five facial components are not difficult to be segmented. The locations and detection confidence output configurations (refer to section 2.3) of these facial components are recorded for future facial components position prediction and pose estimation.

2.2 SSD based region tracking

Then, a fast and robust SSD tracker is proposed to track the mouth-nose region in succeeding frames. Here the mouth-nose region is chosen because it is in the center of the face region which is most slightly affected by the background clutters. So it will be the most reliable sub-facial region under different pose for tracking. In this paper, as a simplified version of SSD tracker proposed by Hager [17], pure translation motion model is adopted from the intuition that a reasonable and mild human motion implies continuous image changes in sequences. Different to Hager's approach, dynamic motion template updates are conducted on each frame to avoid the problem of error accumulation and the reference template is re-initialized whenever the estimated face pose is near frontal.

2.3 Facial component position prediction

According to the above tracked region, facial components are located by facial component detectors. It is done by groups of classifiers to scan all the possible rectangles in the search area, and judge whether the image patch is like a target facial component or not. [15] extended the Ada-boost based object detector by providing a probabilistic-like confidence output for each pixel. But its speed drops as the search area is increased and it requires segmented face region to avoid left/right eye and nose/mouth corner confusion.

In order to reduce the search area and restrict the scales of the target components which need to be scanned, facial geometrical and temporal information are considered. First, mouth corners and nose are detected in the mouth-nose region. Then the positions of left and right eye centers are predicted by affine

transform according to the known mouth corners and nose before detection. Therefore, left/right eye confusion will not exist. The details are described as below.

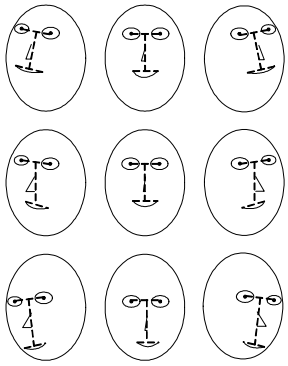

Figure. 2 Planar face model under different poses

Human faces have some special geometric properties, e.g. the bilateral symmetry, which can be used to reduce the noise of facial feature localization [10][11][12]. Anthropometry studies also indicate the discipline of facial features configuration [18]. The line connecting two eye centers and the line connecting two mouth corners are parallel to each other, and these two lines identify the face plane. These four facial feature points and the nose base form a planar face model, which is illustrated in Figure 2. Under orthographic projection model, the parallelism and the length proportion along the same direction are preserved in arbitrary poses, which indicates that the positions of above coplanar facial feature points are transformed under affine model. If the positions of right mouth corner, midpoint of left and right mouth corners and nose base are *A(m, 0)*, *B(0, 0)* and *C(0, n)* respectively which are determined in frontal face during initialization phase. When the face turns to profile pose, these facial feature points are transformed to $A'(x_a, y_a)$, $B'(x_b,y_b)$ and $C'(x_c,y_c)$,

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_1, a_2 \\ a_3, a_4 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} d_x \\ d_y \end{bmatrix} \qquad (1)$$

then the affine transform vector $U(a_1,a_2,a_3,a_4,d_x,d_y)$ is given by $(\frac{x_a - x_b}{m}, \frac{x_c - x_b}{n}, \frac{y_a - y_b}{m}, \frac{y_c - y_b}{n}, x_b, y_b)$ and the positions of eyes can be predicted according to the affine vector and the positions of left/right eye centers initialized on frontal face in section 2.1.

2.4 Facial component detection with confidence output

After applying a facial component detector on each position in the search area predicted in Section 2.3, all the detection confidence outputs are used to weight these positions to determine the final location $\tilde{x}$ and its overall confidence Φ_t, i.e.:

$$\tilde{x} = \frac{\sum_x \Phi_h(x) \cdot x}{\sum_x \Phi_h(x)} \qquad (2)$$

$$\Phi_t = \sum_x \Phi_h(x) \qquad (3)$$

where $\Phi_h(x) = \begin{cases} 0 & \Phi(x) < t \\ \Phi(x) - t & \Phi(x) \geq t \end{cases}$

$\Phi(x)$ is the confidence output on a position. t is a preset threshold which presents the baseline to validate the existence of the target facial component. This weighting strategy considers all the detection results in the search area, so that the final facial component location is accurate and insensitive to noises, comparing with only selecting the position with highest detection confidence.

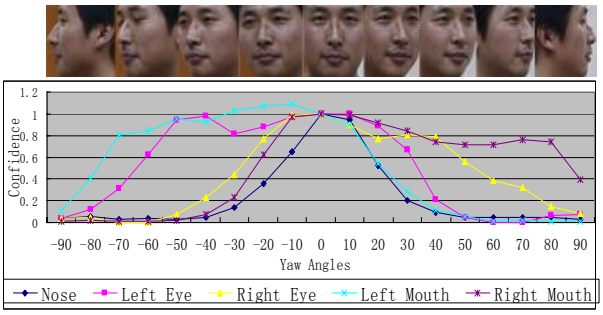

Figure 3. Facial components confidences vs. yaw

3. From coarse to fine face pose estimation

3.1. Coarse pose estimation based on facial component confidences

In Equation (3), Φ_t is intrinsically a metric of the similarities between the detected object and the training samples. When the training facial component samples are cropped in frontal faces, Φ_t characterizes the similarities between the test facial components in a specific pose and facial components in frontal face, which will be higher in near frontal pose and less in profile pose. This property will be used in this section to estimate the pose of a face.

As illustrated in Figure 3, facial component confidences explicitly relate to face pose in two aspects. First, the confidence of the same facial component decreases when a head rotates from frontal to profile pose, which actually indicates the appearance difference to frontal view facial component model. On the other hand, the confidences of corresponding facial components such as left/right mouth corners and left/right eyes also differ, which indicates their

asymmetrical appearance due to the non-frontal pose.

In order to make the above confidence output more stable and reliable, the search regions and the target facial component scales are normalized to guarantee that the confidences of the same facial component under different face poses are comparable. The values of different facial component detection confidences are also normalized regarding to their confidences in initialized frontal face to eliminate the variation of illuminations and identifications.

Finally, a feature vector composed of multiple confidences on eye centers, mouth corners and nose is calculated on each face image. Thus the face pose is estimated by either least square error (LSE) algorithm or other statistical learning algorithms, such as support vector regression (SVR).

The proposed face pose estimation method essentially depends on the facial components confidence distributions instead of the single locations of facial feature points, which is totally different to typical *Geometry-based* approaches. First, the facial component detectors are constructed by learning on large face database of variant identifications, illuminations and expressions; second, the proposed facial component confidence integrates the information from the whole search area, not just one position. These two features make the proposed confidence based pose estimation insensitive to the components localization errors.

3.2. Refine the pose by linear regression

The pose estimation method discussed in section 3.1 is based on general facial component appearance models so that it is appropriate for applications which demand high speed while are not critical of high precision. There will be a slight deviation when the illumination and expression of the faces are dramatically different to the training face samples. This coarse pose estimation will be refined by linear regression algorithm proposed in this section.

3.2.1. Problem setup. To consider the 3D-to-2D transform composed of similarity transformation and orthographic projection, it is assume that $x=(x_1,x_2,x_3)^T$ is the 3D position of one facial component in 3d space, $y=(y_1,y_2)^T$ is its 2d projection and their relationship is

$$\mathbf{y} = s \cdot \mathbf{P} \cdot \mathbf{U}_\gamma \mathbf{U}_\beta \mathbf{U}_\alpha \mathbf{x} + \mathbf{c} + \boldsymbol{\varepsilon} \quad (4)$$

where s denotes the scale factor, P denotes the projection matrix, $U_\alpha, U_\beta, U_\gamma$ denotes the 3D rotation matrix and c denotes the 2D translation vector. The noise ε is assumed to distribute as Gaussian, which is determined in section 3.1 by the facial component detectors.

Supposing that there is a 3D face model with facial feature points $\{x^{(1)},x^{(2)},\ldots,x^{(n)}\}$ and its 2D projection is $\{y^{(1)},y^{(2)},\ldots,y^{(n)}\}$, then from Equation (4), 3D pose estimation is set up by a linear regression problem as,

$$\begin{pmatrix} \mathbf{y}^{(1)} \\ \ldots \\ \mathbf{y}^{(n)} \end{pmatrix} = s \cdot \left[\mathbf{I}_n \otimes \left(\mathbf{P} \cdot \mathbf{U}_\gamma \mathbf{U}_\beta \mathbf{U}_\alpha \right) \right] \cdot \begin{pmatrix} \tilde{\mathbf{x}}^{(1)} \\ \ldots \\ \tilde{\mathbf{x}}^{(n)} \end{pmatrix} + \mathbf{1}_n \otimes \mathbf{c} + \boldsymbol{\varepsilon} \quad (5)$$

where $\otimes$ denotes the Kronecker product of matrices.

3.2.2. Pose estimation by iteratively regression. The coarse pose estimation results from former confidence based method are adopted in the first iteration of the regression. Since β and α have been coarsely estimated, put $\tilde{\mathbf{x}} = \mathbf{U}_\beta \mathbf{U}_\alpha \mathbf{x}$, from Equation (4), the regression model is formulated as

$$\mathbf{Y} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad (6)$$

Where $\boldsymbol{\theta} = \begin{pmatrix} s\cdot\cos\gamma \\ s\cdot\sin\gamma \end{pmatrix}$, $\mathbf{Y} = \mathbf{y}_1 \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \mathbf{y}_2 \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and $\mathbf{A} = \tilde{\mathbf{x}}_1 \otimes \mathbf{I}_2 + \tilde{\mathbf{x}}_2 \otimes \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. The translation item is omitted because $\{x^{(1)},x^{(2)},\ldots,x^{(n)}\}$ are supposed to be centered. So the estimation of the parameters is deduced to

$$\hat{\boldsymbol{\theta}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y} = (\|\tilde{\mathbf{x}}_1\|^2 + \|\tilde{\mathbf{x}}_2\|^2)^{-1} \begin{pmatrix} \tilde{\mathbf{x}}_1^T\mathbf{y}_1 - \tilde{\mathbf{x}}_2^T\mathbf{y}_2 \\ \tilde{\mathbf{x}}_1^T\mathbf{y}_2 + \tilde{\mathbf{x}}_2^T\mathbf{y}_1 \end{pmatrix} \quad (7)$$

After s and γ are determined, β and α will be estimated with the similar method. Then s, γ, β and α are updated iteratively by above process, until the error is below a preset threshold or the max iteration number.

To balance the speed and accuracy, a practical solution is that the coarse pose estimation is performed only when a face is moving fast and the pose is refined by linear regression when the face stops or moves slowly. Its efficiency will be shown in next section with a real-time user attention tracking system.

4. Experiment results and system implementation

4.1. Facial component detection

To test the accuracy of the facial component detection, face images under different yaw poses are collected and tested. The test set (referred as TestSetA later) includes 15 subjects, 20 pictures for each subject covering -90 to +90 degree rotation in yaw direction with interval of 10 degree. Some samples of TestSetA are shown in Figure 3.

Table 1. Facial feature localization error

Components	LEye	REye	LMouth	RMouth	Nose	Average
X Err (pixel)	3.13	2.77	3.30	3.40	7.25	3.97
Y Err (pixel)	1.46	1.76	2.67	2.95	3.52	2.47

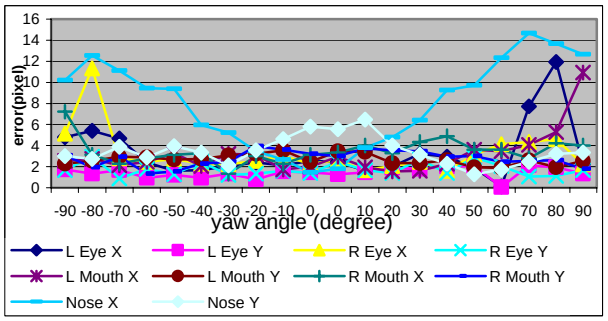

Figure 4. Facial components detection results

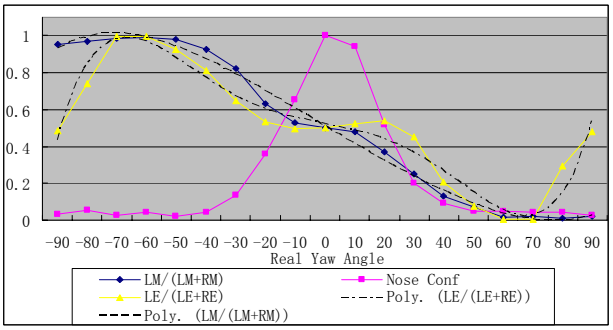

Figure 5. Estimate yaw angle by confidences

On TestSetA of total 300 face images, the average facial component localization error is about 4 pixels in horizontal direction and 3 pixels in vertical direction with standard deviation of 2.98. The face size in these test images is about 150 by 150 pixels, which means that the relative facial component localization error is only about 2.6% and 2% in horizontal and vertical direction respectively.

The test result also indicates the tight relation between the facial component detection errors and the pose of the face. As Figure 4 illustrates, localization errors of left/right eye centers and left/right mouth corners in profile poses are bigger than those in frontal poses, which are caused by the self-occlusion of face. Nose tip localization is also affected by the nosewing in profile face, which leads to bigger horizontal location errors.

4.2. Pose estimation

Pose estimation is conducted on TestSetA based on the facial components detection results in Section 4.1. The first 10 subjects are used for solving the mapping function by LSE and the rest 5 subjects are used for testing. As illustrated in Figure 5, the yaw angle of a face can be recovered by a piecewise linear function

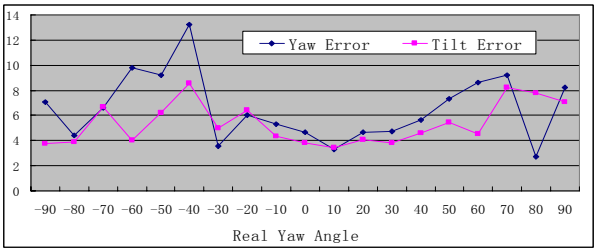

Figure 6. Final pose estimation results

about nose confidence C_{Nose} or polynomial functions about the ratio of left/right mouth/eye confidences. The pose estimation function is as below:

$$Yaw = a_1 Y_{eye} + a_2 Y_{mouth} + a_3 \,\mathrm{sgn}(Y_{eye} + Y_{mouth}) Y_{nose}$$

$$where \quad Y_{eye} = \sum_{i=1}^{K_{eye}} p_i \left(\frac{C_{LeftEye}}{C_{LeftEye} + C_{RightEye}} - \frac{1}{2}\right)^i \qquad (8)$$

$$Y_{mouth} = \sum_{i=1}^{K_{mouth}} q_i \left(\frac{C_{LeftMouth}}{C_{LeftMouth} + C_{RightMouth}} - \frac{1}{2}\right)^i$$

$$Y_{nose} = \begin{cases} 0 & (C_{Nose} < 0.1) \\ k_{nose} C_{Nose} + b_{nose} & (C_{Nose} >= 0.1) \end{cases}$$

Y_{eye}, Y_{mouth} and Y_{nose} are yaw angles estimated by different facial components weighting by. a_1, a_2 and a_3. K_{eye}, K_{mouth}, p_i, q_i, k_{nose} and b_{nose} are constants learnt from training data, in our experiments K_{eye},=K_{mouth}=5, k_{nose}=0.025 and b_{nose}=1. The error of the coarse pose estimation in yaw pose is about 7.75 degree and the errors of the final pose estimation in yaw and tilt angles are 6.54 and 5.35 degree respectively. The errors on different yaw angles are plotted in Figure 6, where the majority of the pose estimation errors on yaw and tilt angles on different views are less than 10 degree. In the real-time pose estimation system implemented in Section 4.3, the pose estimation results are filtered and smoothed by the neighbor frames so that the accuracy and robustness are further improved. The speed of the facial component detection and confidence based pose estimation is about 90 ms/frame. Together with the linear regression algorithm, whose speed is about 2.81 ms/frame (50 iterations), the overall speed is higher than 100 ms/frame on a P4 1.4GHz CPU, which is acceptable for real-time applications.

4.3. Real-time user attention tracking system

Based on the pose estimation results, a real-time user attention tracking system is implemented. It is able to track the attention direction of a computer user in real-time with single camera. The attentive document window is activated automatically when the user turns to it and the mouse/keyboard cursor is put at the original position automatically, so that the user is able to switch among multiple windows without mouse motion. The system achieves real-time (15frame/s) performance on a P4 2GHz CPU and a typical USB web camera with resolution of 320*240 pixel.

Table 2. Pose estimation result comparison

Method	Feature	Accuracy	Speed
Ellipsoidal model, edge density feature point [7]	Wide-Range, person and Illumination-Insensitive	19º for different person, about 10º if initialization for person	200ms~ 300ms on PII 450MHz
PCA + SVR [5]	Kernel-based learning	10º for different person	N/A
KPCA+ SVC[6]	Kernel-based learning, multi-class classification	99% classification accuracy, within 20º,97% within 10º	N/A
Confidence of Feature Point + Regression	Fast, accurate, small change	6.54º in yaw angle and 5.35º in tilt angle if initialization for person	<100 ms on P4 1.4GHz

5. Conclusion and discussion

A coarse-to-fine pose estimation approach has been proposed in this paper. Experiments results have demonstrated that it is more accurate and robust compared to previous works as shown in Table 2. A real-time user attention tracking system has been implemented which enables computer users switching among multiple windows in large scale display without mouse motion. The proposed method have following highlights: (1) A novel pose estimation method based on the confidences of the facial component detectors are proposed which is fast robust; (2) The crucial facial feature localization problem is solved by combining the geometry face model and the confidence output from facial component detectors; (3) Fully automatic and real-time face tracking and pose estimation is achieved in the proposed framework, which is more robust and scalable.

In future works, facial component detection confidence will be utilized for 3D face tracking to provide more accurate pose estimation. Using more facial component detectors and refining the 3D face model during tracking are also considered.

6. References

[1] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, M. Bone, "Face Recognition Vendor Test 2002: Evaluation Report", *FRVT 2002*, March 2003.

[2] R. Gross, J. Shi, and J. Cohn. "Quo vadis face recognition?", *The Third Workshop on Empirical Evaluation Methods in Computer Vision*, 2001.

[3] S. Romdhani and T. Vetter, "Efficient, Robust and Accurate Fitting of a 3D Morphable Model", in *Proc. of IEEE Intl. Conf. on Computer Vision (ICCV)*, 2003, pp59-66.

[4] J. Bruske, E. Abraham-Mumm, J. Pauli, and G. Sommer, "Head pose estimation from facial images with subspace neural networks", In *Proc. of Intl. Neural Network and Brain Conf.*, China, 1998, pp528-531.

[5] Y. Li, S. Gong, and H. Lideldl, "Support vector regression and classification based multi-view face detection and recognition", In *Proc. of IEEE Intl. Conf. on Automatic Face and Gesture Recognition (FG2000)*, France, 2000, pp300-305.

[6] S.Z. Li, Q.D. Fu, L. Gu, B. Scholkopf, Y.M. Cheng, H.J. Zhang, "Kernel Machine Based Learning for Multi-View Face Detection and Pose Estimation", In *Proc. of ICCV*, Vancouver, Canada, 2001.

[7] Y. Wu, K. Toyama and T.S. Huang, "Wide-Range, Person- and Illumination-Insensitive Head Orientation Estimation", *FG2000*, France, 2000, pp183-188.

[8] T. Alter, "3-D pose from 3 points using weak-perspective", In *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 16(8), 1994, pp802-808.

[9] T. Huang, A. Bruckstein, R. Holt, A. Netravali, "Uniqueness of 3D Pose under Weak Perspective: A Geometrical Proof", *PAMI*, 1995 vol.17, pp1220-1221.

[10] Z. Liu, Z. Zhang, "Robust Head Motion Computation by Taking Advantage of Physical Properties", *IEEE Workshop on Human Motion*, Austin, USA, Dec. 2000.

[11] Y. Liu, K. Schmidt, J. Cohn, R.L. Weaver, "Facial Asymmetry Quantification for Expression Invariant Human Identification", *In Proc. of FG2002*.

[12] Y. Liu, R.L. Weaver, K. Schmidt, N. Serban, J. Cohn, "Facial Asymmetry: A New Biometric", *Technical Report CMU-RI-TR-01-23*, Robotics Institute, Carnegie Mellon University, Aug., 2001.

[13] Y. Zhou, L. Gu, H.J. Zhang. "Bayesian Tangent Shape Model - Estimating Shape and Pose Parameters via Bayesian Inference", In *Proc. of IEEE Conf. on Pattern Recognition (CVPR)*, 2003.

[14] S.C. Yan, C, Liu, S.Z. Li, H.J. Zhang, H. Shum, "Face Alignment Using Texture-constrained Active Shape Models", In *Proc. of European Conf. on Computer Vision, Workshop on Generative-Model-Based Vision*, 2002.

[15] L.B. Chen, L. Zhang, L. Zhu, M.J. Li, H.J. Zhang, "A Novel Facial Feature Point Localization Algorithm Using Probabilistic-like Output", In *Proc. of Asian Conf. on Computer Vision* 2004

[16] P. Viola and M. J. Jones, "Robust real-time object detection", *Technical report*, COMPAQ Cambridge Research Laboratory, Cambridge, MA, February 2001.

[17] G. Hager and P. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination", *PAMI*, 98 20(10), pp1025-1039.

[18] K. Lam, H. Yan, "An Analytic-to-Holistic Approach for Face Recognition Based on a Single Frontal View", *PAMI*, 98 20(7), pp673-686