

Multi-View Face Alignment Using 3D Shape Model for View Estimation

Yanchao Su¹, Haizhou Ai¹, Shihong Lao²

¹Computer Science and Technology Department, Tsinghua University

²Core Technology Center, Omron Corporation
ahz@mail.tsinghua.edu.cn

Abstract. For multi-view face alignment (MVFA), the non-linear variation of shape and texture, and the self-occlusion of facial feature points caused by view change are the two major difficulties. The state-of-the-art MVFA methods are essentially view-based approaches in which views are divided into several categories such as frontal, half profile, full profile etc. and each of them has its own model in MVFA. Therefore the view estimation problem becomes a critical step in MVFA. In this paper, a MVFA method using 3D face shape model for view estimation is presented in which the 3D shape model is used to estimate the pose of the face thereby selecting its model and indicating its self-occluded points. Experiments on different datasets are reported to show the improvement over previous works.

Keywords: Active Shape Model, face alignment, 3D face model

1 Introduction

Automatically locating facial feature points on face images, i.e. face alignment (FA) is a critical task in many face related computer vision areas such as 3D face labeling, expression analysis and face recognition. For face alignment, there are two fundamental approaches, Active Shape Model (ASM) [1] and Active Appearance Model (AAM) [2]. Many variations of these two methods have been developed to improve their robustness and accuracy [3-5]. While for frontal FA there is already some robust algorithm can be used in practice [4], for MVFA where face is with large view change, it remains a challenging problem since both the shape and the texture of face in images change dramatically when the view changes.

In the literature of MVFA, non-linear modeling method such as Gaussian Mixture Model [6], kernel PCA [7], Bayesian Mixture Model with learning visibility of label points [8] and view-based methods such as view based DAM [9] and view-based ASM [10] are developed which are mainly 2D approaches with no appealing to 3D face information. Due to the intrinsic difficulties caused by face appearance changes in 2D face images of a 3D face, MVFA is still not a solved problem.

The state-of-the-art MVFA methods are essentially view-based approaches in which views are divided into several categories such as frontal, half profile, full profile etc. and each of them has its own shape and texture models. Since the texture model used in local search of each label point of a particular shape model depends on

its view category, these methods are very sensitive to the estimation of the view category. When the initial view is not predicted correctly, the results of local search become unreliable. And if the estimation of shape parameter does not deal with the potential outliers, ASM approach will fail.

In the original view-based ASM method [10] a non-linear optimization method for model selection is used in which each feature point is weighed dynamically so that only the feature points that are consistent with the shape model will have large weights, while the effect of outliers will be eliminated. Since this method does not completely depend on the local search of each label points, it is more robust against the initial view and cluttered background.

View based methods switch between different models of different views to cover the non-linear space of multi-view faces, thus, the selection of models, in other words, the estimation of view is a critical step in the algorithm. Although the overlapped definition of view ranges can mitigate the error caused by improper initialization of view, the automatic view estimation in the alignment procedure is still an important problem to be solved.

There are other MVFA approaches using 3D face model. In [11], view-based local texture models and a sparse 3D shape model which are trained using synthesized faces are integrated in an ASM-like framework to align faces with view changes. In [12], a parameterized 3D deformable face model is used to help with view based ASM but building its 3D face model is a very tough work.

In this paper, we combine a view-based ASM and a simple 3D face shape model built on 500 3D-scanned faces [13] to build a fully automatic MVFA system. Initialized by a multi-view face detector [14], we first use view based local texture model to local search the feature points around the initial shape [10], then a 3D face shape is reconstructed from those points using the 3D face shape model. According to the reconstructed 3D shape, we can get its view information from which self-occluded points can be indicated, and then the 2D shape model of this view is adopted to refine the observed non-occluded shape by non-linear parameter estimation.

2 View-Based Active Shape Model

In the case of MVFA, the shape and texture of faces in images change dramatically when the view changes. A single PCA model can only represent face shapes with limited view change due to the non-linear change of face shape. And further textures around the label points with large view changes are also hard to be characterized in a single model. So as in [10] we divide views into 7 categories and for each view we train a set of local texture models and a shape model. Therefore in MVFA, a face shape is represented by a PCA model of the view \mathcal{V} it belongs to:

$$S = T_q \left(U_v \cdot p + \bar{S}_v \right) \quad (1)$$

So the objective of MVFA is to find the best PCA parameter p and pose q with some view \mathcal{V} :



Fig. 1. Mean shapes of 7 view categories

$$(p, q) = \arg \max P(S|v) \prod P_i(I|(x_i, y_i), v) \quad (2)$$

where $P(S|v) \sim N(\bar{S}_v, \text{diag}(\lambda_{1,v}, \dots, \lambda_{m,v}))$ is the shape prior of specified view v ($\lambda_{i,v}$ is the i -th eigenvalue of the covariance matrix of shape samples with view v). And $P_i(I|(x_i, y_i), v)$ is the probability of the point (x_i, y_i) to be the i -th label point, which is determined by local texture model. The local texture model of each view is trained using Haar-like feature based boosted classifier [4] which given a texture patch can output the likelihood of this patch to be around the i -th label point.

The whole alignment procedure is as follow:

1. Given an image I , the bounding box and the estimated view v_0 is provided by the face detection module. And the initial shape S_0 is estimated by fitting the mean shape of v_0 into the bounding box of face. See Figure 1 for mean shape illustration.
2. For each label point, search locally around its current position for the best displacement (x_i^*, y_i^*) with the largest likelihood using the local texture models of current view.
3. Parameter estimation: for each view, estimate the parameter $p_{v'}$ and $q_{v'}$ using non-linear parameter estimation and then find the best view v' and its corresponding parameter p' and q' with the minimum reconstruction error of the shape.
4. Set the new shape $S = T_{q'}(U_{v'} \cdot p' + \bar{S}_{v'})$, and current view $v = v'$
5. Iterates from step 2 until the shape S converged.

In the optimization of MVFA, the proper selection of the hidden view v , is a critical step which will severely affect its accuracy and robustness. So we have to develop a robust pose estimation method to select a proper view when an inaccurate initial view is given by the face detector.

3. 3D face shape model

While a 2D face model suffers when view changes, a 3D face model can easily overcome this obstacle. The 2D ASM deals both the intrinsic change (caused by the change of expression and different person) and the extrinsic change (caused by image projection) with a single linear model; while the 3D shape model reflects only the intrinsic change.

Similar as the 2D face shape model, a 3D face shape can be denoted by a list of 3D coordinates $S_{3d} = [x_1, y_1, z_1, \dots, x_n, y_n, z_n]$ and here we use $n=88$ points in 3D-scanned faces

which correspond to the 88 face alignment feature points in 2D face images. And our 3D face shape model is a PCA point distribution model of the 88 feature points built on a 3D-scanned BJUT-3D Face Database [13].

The first step to construct the 3D face shape model is to achieve the 3D locations of the 88 feature points in each 3D-scanned face as follows:

Each 3D-scanned face is rendered at various views through orthogonal projection, and then the 2D ASM is employed to obtain 2D feature point locations in the rendered images. And the corresponding 3D location of each feature point can be achieved by the following method:

Given a projection matrix (in fact, it is a coordinate transformation matrix) $P_i=(U_i^T V_i^T W_i^T)$, the projected coordinate of the i -th feature point $X=S_{3d,i}$ is $(x_i, y_i, z_i)^T=P_i S_{3d,i}$. The 2D ASM gives an observation of x_i as \tilde{x}_i . The depth channel of the rendered face gives an observation of Z_i as $\tilde{z}_i = \phi(\tilde{x}_i, \tilde{y}_i)$. We assume the error p.d.f. of the 2D ASM is Gaussian:

$$\begin{aligned}\tilde{x}_i &\sim N(x_i, \sigma_i^2) \\ \tilde{y}_i &\sim N(y_i, \sigma_i^2)\end{aligned}\quad (3)$$

For simplicity we assume the errors on different axes distribute independently. Thus the error of \tilde{z}_i can be derived as

$$\begin{aligned}\tilde{z}_i &= \phi(\tilde{x}_i, \tilde{y}_i) \\ &= \phi(x_i + \delta x_i, y_i + \delta y_i) \\ &\approx \phi(x_i, y_i) + \delta x_i \phi_x(x_i, y_i) + \delta y_i \phi_y(x_i, y_i) \\ &= z_i + \delta x_i \phi_x(x_i, y_i) + \delta y_i \phi_y(x_i, y_i) \\ \tilde{z}_i &\sim N(z_i, \theta_i^2)\end{aligned}\quad (4)$$

where $\theta_i^2 = (\phi_x^2(x_i, y_i) + \phi_y^2(x_i, y_i)) \sigma_i^2$

Rotating the 3D face and doing 2D ASM on the rendered faces give several sets of projection matrix and observed coordinates $(P_i, \tilde{x}_i, \tilde{y}_i, \tilde{z}_i)$. The log joint likelihood is:

$$\begin{aligned}L(X) &= \log(P((\tilde{x}_i, \tilde{y}_i, \tilde{z}_i), i=1 \dots n | X)) \\ &= \sum_i (\log P(\tilde{x}_i | X) + \log P(\tilde{y}_i | X) + \log P(\tilde{z}_i | X)) \\ &= \lambda - \sum_i \left(\frac{(\tilde{x}_i - U_i X)^2}{2\sigma_i^2} + \frac{(\tilde{y}_i - V_i X)^2}{2\sigma_i^2} + \frac{(\tilde{z}_i - W_i X)^2}{2\theta_i^2} \right)\end{aligned}\quad (5)$$

where λ is constant with respect to X .

The maximum likelihood estimation can be obtained analytically by letting the derivative to be zero.

$$X_{ML} = \arg \max_x L(X) \quad (6)$$

$$\nabla X = 0$$

In practice the 2D ASM is not always accurate. So we use RANSAC algorithm for robust estimation of X .

After achieving the 3D locations of 88 feature points for each of the 500 3D-scanned faces in BJUT-3D Face Database [13], we select 389 3D face shapes among them which are good estimations to build the 3D point distribution model using PCA as the 3D face shape model.

$$S_{3d} = U \cdot p + \bar{S}_{3d} \quad (7)$$

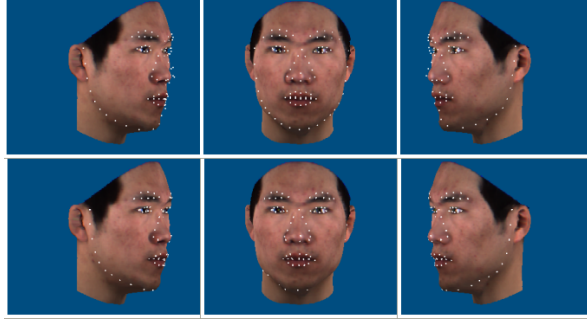


Fig. 2. Achieving 3D locations of feature points. First row shows 2D ASM results on different views. Second row shows reconstructed 3D position of feature points on those views.

4. View estimation using 3D shape model

Given an input image, suppose the orthogonal projection holds, the 2D shape of the face denoted as S_{2d} in the image should be:

$$S_{2d,i} = M \cdot P \cdot S_{3d,i} + t \quad (8)$$

Each label point $S_{3d,i}$, which is determined by the PCA parameter p in equation (7), is first transformed (scaled and rotated) by the transform matrix P , then projected into the image plane orthogonally with translation t . Where P is an orthogonal projection matrix and $M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ is the projection matrix.

According to equation (8), given a candidate 2D face shape S_{2d} , we could get the PCA parameter p and the pose information (s, R, t) using minimum square error estimation:

$$(P, t, p) = \arg \min \sum_i \left\| M \cdot P \cdot (U \cdot p + \bar{S}_{3d,i}) + t - S_{2d,i} \right\|^2 \quad (9)$$

We first reform the object function by denoting $A=M \cdot P$, so that:

$$(A, t, p) = \arg \min \sum_i \left\| A \cdot (U \cdot p + \bar{S}_{3d,i}) + t - S_{2d,i} \right\|^2 \quad (10)$$

where A has the constraint:

$$a_{1,1}^2 + a_{1,2}^2 + a_{1,3}^2 = a_{2,1}^2 + a_{2,2}^2 + a_{2,3}^2, \quad a_{1,1}a_{2,1} + a_{1,2}a_{2,2} + a_{1,3}a_{2,3} = 0 \quad (11)$$

which makes A an orthogonal projection.

We can solve the above optimization problem by optimizing pose parameter (A, t) and shape parameter p alternatively in an iteration procedure as follows:

1. Given p , solve (A, t)

$$(A, t) = \arg \min \sum_i \left\| A (U \cdot p + \bar{S}_{3d,i}) + t - S_{2d,i} \right\|^2 \quad (12)$$

An affine projection (A', t) can be estimated analytically and then we can get an orthogonal projection A by optimizing the object function using gradient descend initiated with the affine projection.

2. Given (A, t) , solve p

$$p = \arg \min \sum_i \left\| A (U \cdot p + \bar{S}_{3d,i}) + t - S_{2d,i} \right\|^2 \quad (13)$$

It is a linear MSE and the solution goes straight forward.

Step 1 and Step 2 are iterated until the convergence of the reconstruction error.

5. Automatic MVFA

Given a face image, we initialize our algorithm by applying multi-view face detection [14] which provides a bounding box and a roll angle of the face. The roll angle corresponds to 5 view categories in $\{-90^\circ, -45^\circ, 0^\circ, 45^\circ, 90^\circ\}$. We select the initial view according to this angle and then compute the initial 2D shape by rotating and scaling the mean shape of initial view to fit the bounding box of face. Then the algorithm goes iteratively as follow:

1. Local Search: For the i -th label point, compute the likelihood $P(I(x_i, y_i), v)$ using the local texture model of current view on every point around the current location of label point i , then select the best candidates $\{(x_i^*, y_i^*)\}$ with the largest likelihood as the new location. The observation shape is $S_{2d}^* = \{(x_i^*, y_i^*)\}$.

2. Pose Estimation using the 3D face shape model: use the observation 2D shape and the 3D shape model to estimate the pose parameter (A, t) and the shape parameter p of the 3D face model. Then compute the roll angle according to the projection matrix A and select current view. At the same time we can indicate the self-occluded label points by the reconstructed 3D shape.

3. 2D parameter estimation: estimate the shape and pose parameter using 2D shape model of current view. Given the observed shape S_{2d}^* and the visibilities of each label point, the new shape is reconstructed by minimizing the weighted

reconstruction error of visible points. The dynamic weighting method used in [10] is still adopted in our algorithm to improve robustness.

Here is the flow chart of our automatic MVFA system.

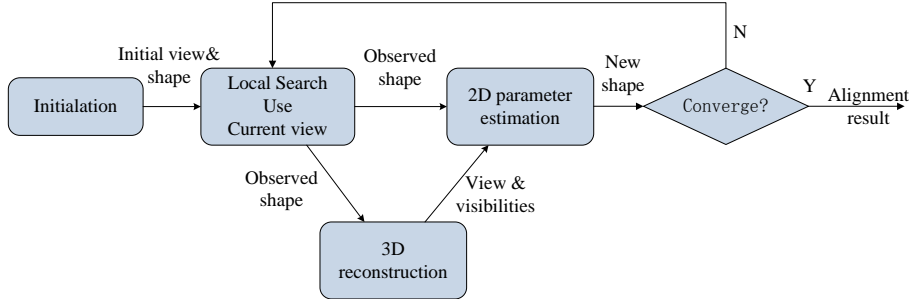


Fig. 3. Flowchart of MVFA

The whole alignment procedure is shown in Figure 4.

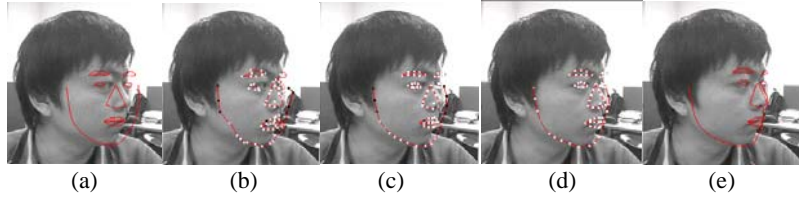


Fig. 4. Illustration of alignment procedure ((a) The algorithm is initialized by the mean shape of current view. (b) The observed shape is got by local search using local texture model. (c) With the observed shape, a 3D shape is reconstructed using 3D shape model and the pose is estimated. (d) The 2D shape is reconstructed from the observed shape. (e) The final shape when the iteration converged.)

6. Experiments

6.1 Training

A multi-view face database including totally 1800 images which are taken by a camera array with poses in the set $\{-90^\circ, -75^\circ, -60^\circ, -45^\circ, -30^\circ, 0^\circ\}$ is set up. The face images are about 250 by 250 in size each of which are manually labeled with 88 label points. 1500 images are used in training and the other 300 are used in testing. 4 ASMs of corresponding views shown in table 1 are trained using Haar-like feature based boosted classifiers which distinguish the textures around a feature point from the textures far away from the feature point [10] (here the view 5-7 are omitted since they are the mirrors of the view 2-4 which can use their mirrored models). Notice that the angle ranges of different views have overlaps in order to make each model more robust to view selection.

Table. 1 Training models and roll angles

View	1	2	3	4
Angle	Frontal	0 to 45	30 to 60	60 to 90

6.2 View estimation

View estimation results are tested on the 1500 training images. Table 2 gives the comparison results between the 3D approach and the 2D view-based approach. It can be seen that the 3D method can apparently improve the view estimation accuracy especially for those views with large off-image-plane (roll) angles which are very critical in MVFA since face alignment for faces of non-frontal views are much more sensitive to view selection.

Table. 2 Comparison between the 3D approach and the 2D view-based approach

View	0	1	2	3
3D method	95%	93%	92%	95%
2D method	93%	90%	87%	85%

6.3 MVFA

On the 300 testing images, the performance of MVFA is measured by the average point-to-point errors between alignment result and the ground truth and is shown in Figure 5. It can be seen that the proposed approach outperforms the traditional view-based ASM algorithm [10]. In average, MVFA takes about 304ms to align a face.

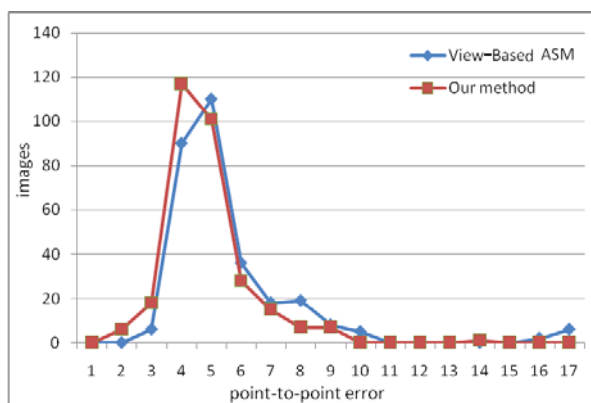


Fig. 5. Error distribution of alignment results.

We also tested our method on the CMU-PIE database. Some results are shown in Figure 6. Since there are no ground truth data, we can only subjectively judge the correctness of alignment on a subset of CMU-PIE database. Among all the 1145 face images from the c02, c37, c27, c11, c14 view categories, our algorithm achieved 86.7% in correct rate, while the original method [10] can only achieve 74.5%.

Additional tests have also been taken on the Labeled faces in the wild database [16], which contains multi-view faces in unconstrained environments. Our method can deal with these faces rather well even though our training images are taken in constrained environments and do not cover such large variations in pose, illumination, background, focus and expression. See figure 7 for some results.



Fig. 6. Additional results on CMU-PIE database



Fig. 7. Additional results on Labeled faces in the wild database

7 Conclusion

In this paper, we presented an automatic MVFA framework by integrate 2D view based ASM with a 3D face shape model. Alignment is done in view-based ASM manner while during the iterations, the selection of models, in other words, the view estimation, is done using 3D face shape model. In addition, the 3D reconstructed shape is used to indicate invisible label points that can further improve the accuracy and robustness of the 2D view-based ASM method. Experiments show that view estimation using 3D model can help the view-based ASM method in both accuracy and robustness. Our future work will focus on extending the proposed method to more challenging datasets such as the Labeled faces in the wild database and consumer images over the internet.

Acknowledgement

This work is supported by National Science Foundation of China under grant No.60673107, and it is also supported by a grant from Omron Corporation.

References

1. A. Hill, T.F. Cootes, and C.J. Taylor, Active shape models and the shape approximation problem, BMVC 1995.
2. T.F. Cootes, G.J. Edwards, and C.J. Taylor, Active appearance models, IEEE Transactions on pattern analysis and machine intelligence, vol. 23, NO. 6, June 2001.
3. F. Jiao, S.Z. Li, et.al, Face alignment using statistical models and wavelet features, CVPR 2003.
4. L. Zhang, H. Ai, et.al, Robust Face Alignment Based on Local Texture Classifiers, ICIP 2005.
5. A.U. Batur, M.H. Hayes. A Novel Convergence for Active Appearance Models. CVPR 2003.
6. T.F. Cootes and C.J. Taylor. A mixture model for representing shape variation. BMVC 1997.
7. S. Romdhani, S. Gong, and A. Psarrou. A multi-view non-linear active shape model using kernel PCA. BMVC 1999.
8. Y. Zhou, W. Zhang, et.al, A Bayesian Mixture Model for Multi-view Face Alignment, CVPR 2005
9. S.Z. Li, S.C. Yan, et.al, Multi-view face alignment using direct appearance models, AFG 2002
10. L. Zhang, H. Ai, Multi-View Active Shape Model with Robust Parameter Estimation, ICPR 2006.
11. L. Gu and T. Kanade, 3D Alignment of Face in a Single Image, CVPR 2006.
12. C. Vogler, Z.G. Li, A. Kanaujia, The Best of Both Worlds: Combining 3D Deformable Models with Active Shape Models, ICCV 2007
13. The BJUT-3D Large-Scale Chinese Face Database. Technical Report No ISKL-TR-05-FMFR-001. Multimedia and Intelligent Software Technology Beijing Municipal Key Laboratory, Beijing University of Technology, 2005.
14. C. Huang, H. Ai, et.al, High Performance Rotation Invariant Multiview Face Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.29, No.4, pp. 671-686, APRIL 2007.
15. T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression (PIE) database of human faces. the roboticsinstitute, Carnegie Mellon University. Technical report, 2001.
16. Huang, G.B., Ramesh, M., Berg, T., Miller, E.L.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report, 25(12):07-49 (October 2007).