Face Alignment Using Statistical Models and Wavelet Features

Feng Jiao[1*], Stan Li[2], Heung-Yeung Shum[2], Dale Schuurmans[1]
[1]Department of Computer Science, University of Waterloo
Waterloo, Canada
[2]Microsoft Research Asia
Beijing, China
{fjiao@math.uwaterloo.ca}

Abstract

Active Shape Model (ASM) is a powerful statistical tool for face alignment by shape. However, it can suffer from changes in illumination and facial expression changes, and local minima in optimization. In this paper, we present a method, W-ASM, in which Gabor wavelet features are used for modeling local image structure. The magnitude and phase of Gabor features contain rich information about the local structural features of face images to be aligned, and provide accurate guidance for search. To a large extent, this repairs defects in gray scale based search. An E-M algorithm is used to model the Gabor feature distribution, and a coarse-to-fine grained search is used to position local features in the image. Experimental results demonstrate the ability of W-ASM to accurately align and locate facial features.

1. Introduction

Accurate face alignment is important for extracting good facial features, which in turn is important for achieving success in applications such as face recognition, expression analysis and face animation. Extensive research has been conducted on image feature alignment over the past 20 years. For example, Kass *et al* [1] introduced Active Contour Models, an energy minimization approach for shape alignment. Kirby and Sirovich [2] described statistical modeling of grey-level appearance but did not address face variability. Wiskott *et al* [3] used Gabor Wavelets to generate a data structure named the Elastic Bunch Graph to locate facial features. This latter approach can tolerate a certain degree of pose and expression change, and has proven to be very useful. It searches for facial points on the whole image and uses the distortion of the graph to adjust the feature points. Unfortunately, this procedure is time-consuming and requires significant computation.

Active Shape Models (ASM) and Active Appearance Models (AAM), proposed by Cootes *et al* [4][5], are two popular shape and appearance models for object localization. They have been developed and improved for many years.

In ASM [4], the local appearance model, which represents the local statistics around each landmark, allows for an efficient search to be conducted to find the best candidate point for each landmark. The solution space is constrained by properly training a global shape model. Based on modeling local features accurately, ASM obtains good results in shape localization. AAM [5] combines constraints on both shape and texture in its characterization of facial appearance. In the context of this paper, texture means the intensity patch contained in the face shape after warping to the mean face shape. There are two linear mappings assumed for optimization: from appearance variation to texture variation, and from texture variation to position variation. The shape is extracted by minimizing the texture reconstruction error. According to the different optimization criteria, ASM performs more accurate shape localization while AAM gives a better match to image texture. On the other hand, ASM tends to get stuck in local minima, depending on initialization. AAM is sensitive to illumination, particularly if the lighting during testing is significantly different from the lighting during training. In addition, training an AAM model is time consuming.

In this paper, we present an improved ASM method, called W-ASM, in which Gabor-Wavelet features are used to model local structures of the image. The magnitude and phase of Gabor features contain rich information about the local structure of the face to be aligned, and provide accurate guidance for the search. This, to a large extent, repairs defects in gray value based search. An E-M algorithm is used to model the Gabor feature distribution, and a coarse-to-fine approach is used to search for positions for the local points. Compared with the original method used in ASM, W-ASM can achieve more accurate results. Compared with the Elastic Bunch Graph Matching method, by exploiting a statistical model to restrict shape variation, computation is reduced because

* The work described in this paper was performed at Microsoft Research Asia in Beijing.

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

we can use the prior model to direct the search more effectively, rather than search the whole image. Experimental results demonstrate that W-ASM achieves better results than ASMs .

The rest of the paper is organized as follows: The original ASM algorithm is briefly described in Section 2. In Section 3, we present our Gabor based representation for local structures of shape, and an E-M method for computing a more efficient Gabor representation. Our method of search is presented in Section 4. Experimental results are presented in Section 5, and conclusions are drawn in Section 6.

2. Overview of the ASM Algorithm

2.1. Statistical Shape Models

We describe briefly the statistical shape models used to represent deformable objects.

The ASM technique relies upon each object or image structure being represented by a set of points. The points can represent the boundary, internal features, or even external features, such as the center of a concave section of boundary. Given a set of training images for a given object, points are manually placed in the same location on the object in each image. By examining the statistics of the positions of the labeled points a "Point Distribution Model" is derived. The model gives the average positions of the points, and has a number of parameters that control the main variations found in the training set.

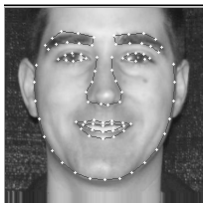

Figure 1. Labeled image with 87 landmarks

The points from each image are represented as a vector x and aligned to a common co-ordinate frame. Principle Component Analysis [2] is applied to the aligned shape vector

$$x = \bar{x} + P\,b \tag{1}$$

where $\bar{x}$ is the mean shape vector, P is a set of orthogonal models of shape variation and b is a vector of shape parameters.

The vector b defines a set of parameters for a deformable model. By varying the elements of b we can vary the shape using Equation (1). By applying bounds to the value of parameter b we ensure that the generated shapes are similar to those in the original training set.

The ASM search procedure is an iterative procedure. On each iteration it uses the local appearance model to find a new shape and then updates the model parameters to best fit the new search shape [4].

2.2. Local Appearance Models

The local appearance models, which describe local image features around each landmark, are modeled as the first derivative of the sample profiles perpendicular to the landmark contour [4].

It is assumed that the local models are distributed as a multivariate Gaussian. For the *jth* landmark, we can derive the mean profile $\bar{g}_j$ and the sample covariance matrix S_j from the *jth* profile examples directly. The quality of fitting a feature vector g_s at test image location s to the *jth* model is given by calculating the Mahalanobis distance from the feature vector to the *jth* model mean.

$$f_j(g_s) = \left(g_s - \bar{g}_j\right)^t S_j^{-1} \left(g_s - \bar{g}_j\right) \tag{2}$$

At the current position s, when searching points, the local appearance models find the best candidate in the neighborhood of the search point, by minimizing $f_j(g_s)$, which is equivalent to maximizing the probability that g_s comes from the distribution.

Using local appearance models leads to fast convergence to the local image evidence. However, due to the variation of the illumination and image quality, often a feature point cannot accurately located. As a consequence, ASM tends to get stuck at local minima, depending on initialization.

3. Modeling Local Features Using Gabor Wavelets

3.1. Gabor Wavelet Representation of Local Features

The use of 2D Gabor wavelet representations in computer vision was pioneered by Daugman in the 1980's [6]. The Gabor wavelet representation allows for a description of spatial frequency structure in the image while preserving information about spatial relations.

A complex-valued 2D Gabor function is a plane wave restricted by a Gaussian envelope:

$$\varphi_j(\vec{x}) = k_j^2 \exp\left(-\frac{k_j^2 x^2}{2\sigma^2}\right)\left[\exp\left(i\,\vec{k}_j\,\vec{x}\right) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \tag{3}$$

$$\vec{k}_j = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_v \cos\phi_u \\ k_v \sin\phi_u \end{pmatrix} \qquad k_v = 2^{-\frac{v+2}{2}}\pi$$

Commonly, 5 frequencies and 8 orientations are used

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

$$\phi_u = u\frac{\pi}{8}, j = u + 8v, v = 0, \cdots 4, u = 0 \cdots 7$$

The second term in the square bracket of Eq. (3) makes the kernels *DC-free,* i.e. the integral $\int \varphi_j(\vec{x}) d^2\vec{x}$ vanishes, which renders the filters insensitive to the overall level of illumination.

Figure 2 shows the 40 standard Gabor kernels.

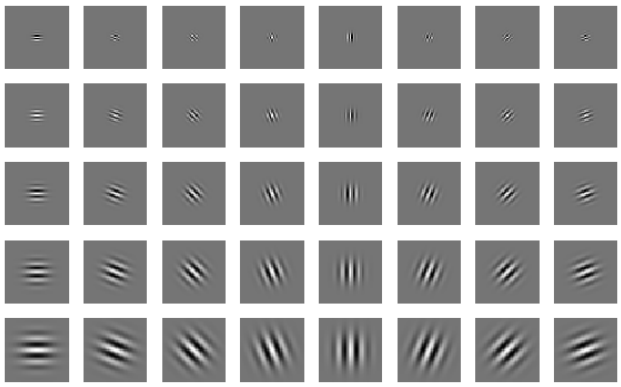

Figure 2. 40 Gabor kernels used in this paper

In an image, for a given pixel $\vec{x}$ with gray level $L(\vec{x})$, the convolution can be defined as

$$J_j\left(\vec{x}\right) = \int L\left(\vec{x}'\right)\varphi_j\left(\vec{x} - \vec{x}'\right)d^2\vec{x}' \quad (4)$$

When all 40 kernels are used, 40 complex coefficients are determined. We refer to this set of 40 coefficients as a jet, which is used to represent the local features. Specifically, a jet is the set of convolution coefficients for kernels of different orientations and frequencies at one image pixel. It can be expressed as $J_j = a_j \exp(i\phi_j)$, where the magnitudes $a_j(\vec{x})$ vary slowly with position, and the phases $\phi_j(\vec{x})$ rotate at a rate approximately determined by the spatial frequency of the kernel.

Two similarity functions are applied. One is a phase-insensitive similarity function:

$$S_a(J, J') = \frac{\sum_j a_j a_j'}{\sqrt{\sum_j a_j^2 \sum_j a_j'^2}} \quad (5)$$

which varies smoothly with the change of the position. The other is a phase-sensitive similarity function:

$$S_\phi(J, J') = \frac{\sum_j a_j a_j' \cos\left(\phi_j - \phi_j' - \vec{d}\,\vec{k}_j\right)}{\sqrt{\sum_j a_j^2 \sum_j a_j'^2}} \quad (6)$$

which changes rapidly with location, and provides a means for accurately localizing jets in an image. Assuming that two jets J and J' refer to object locations with small relative displacement $\vec{d}$, a standard method [7][8] is used to estimate $\vec{d}$:

By expanding Equation (6) in its Taylor form, we obtain

$$S_\phi(J, J') \approx \frac{\sum_j a_j a_j' \left[1 - 0.5\left(\Phi_j - \Phi_j' - \vec{d}\,\vec{k}_j\right)^2\right]}{\sqrt{\sum_j a_j^2 \sum_j a_j'^2}} \quad (7)$$

Setting $\frac{\partial}{\partial d_x} S_\Phi = \frac{\partial}{\partial d_y} S_\Phi = 0$ and solving for $\vec{d}$ leads to

$$\vec{d}(J, J') = \begin{pmatrix} d_x \\ d_y \end{pmatrix} = \frac{1}{\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx}} \times \begin{pmatrix} \Gamma_{xx} & -\Gamma_{xx} \\ -\Gamma_{xx} & \Gamma_{xx} \end{pmatrix} \begin{pmatrix} \Phi_x \\ \Phi_y \end{pmatrix} \quad (8)$$

where if $\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx} \neq 0$, then

$$\Phi_x = \sum\nolimits_j a_j a_j' k_{jx}\left(\Phi_j - \Phi_j'\right)$$
$$\Gamma_{xy} = \sum\nolimits_j a_j a_j' k_{jx} k_{jy}$$

for $\Phi_y, \Gamma_{xx}, \Gamma_{yx}, \Gamma_{yy}$ defined accordingly.

Through this equation, we can estimate the displacement between two jets taken from object locations sufficiently close that their Gabor Kernels are highly overlapped. This approach can estimate displacements up to half the wavelength of the highest frequency kernel, which will be 8 pixels when using the lowest frequency kernels (where v=4), and 2 pixels when using the highest frequency kernels (where v=0). Wiskott uses this method in the algorithm of the Elastic Bunch Graph Matching for face recognition.

This representation is often favored for its biological relevance and technical properties. The Gabor kernels resemble the receptive field profiles of simple cells in the visual pathway. They are localized in both the space and frequency domains and achieve the lower bound of the space-bandwidth product as specified by the uncertainty principle [9].

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

3.2. Modeling Local Features Using the E-M Algorithm

For the labeled training set, the jet of each point in each image, J_{pi}, is calculated, where $i = 1 \cdots N$, N is the number of training images, $p = 1 \cdots M$, and M is the number of the landmark points (87 in our system). We use the jets in the training set to model the local features.

One simple way to model local features is to calculate the mean jet of each landmark in all training images.

$$\bar{J}_p = \frac{1}{N} \sum_{i=1}^{N} J_{pi} \tag{9}$$

Due to changes in background and illumination, the jet values in the same position may vary considerably. For example, sometimes women have long hair which covers the contour of their faces, while men often have shorter hair, which makes the respective jet values totally different. To use the mean value of all jets to represent the entire set of jet values may lead to error.

Here we assume the jet values of each landmark are distributed as a multivariate Gaussian. In order to model the distribution of jets, we use the Expectation-Maximization (EM) algorithm [10] to determine the maximum likelihood parameters for a Gaussian mixture.

For each landmark *i*, we obtain the jets J_{pi}. The distribution of Gabor jets for one landmark are then modeled by the pdf:

$$P\left(J_{pi} \mid W_s, c_s, \mu_{sk}, \Sigma_{sk}, \pi_{sk}\right) = \sum_{k=1}^{c_s} \pi_{sk} G\left(J_{pi} \mid k\right)$$

$$= \left(2\pi\right)^{-1} \sum_{k=1}^{c_s} \pi_{sk} \left|\Sigma_{sk}\right|^{-1/2} \exp\left[-\lambda_{sk}^2 / 2\right] \tag{10}$$

where $\lambda_{sk}^2 = \left[J_{pi} - \mu_s\right]^T \Sigma_s^{-1} \left[J_{pi} - \mu_s\right]$, c_s is the number of Gaussian components, and π_{sk} is the prior probability that the data J_{pi} was generated by the Gaussian component *k*, which satisfies the normalization constraint $\sum_{k=1}^{c_s} \pi_{sk} = 1$. The Gaussian densities $G\left(x_s\left(x, y, z\right) \mid k\right)$ have means μ_{sk} and covariance matrices Σ_{sk}.

The parameters of a Gaussian mixture density can be estimated by maximizing the likelihood function through an iterative procedure known as the Expectation-Maximization (EM) algorithm. The EM algorithm is used for finding maximum likelihood parameter estimates when there is missing or incomplete data. We estimate values to fill in for the incomplete data (the "E-Step"), compute the maximum likelihood parameter estimates using this data (the "M-Step"), and repeat until a suitable stopping criterion is reached.

The E-Step consists of evaluating posterior probabilities of the *kth* Gaussian kernel given the jet for each mixture component. First the posterior probability in each Gaussian component is calculated

$$P_{l,k}\left(k \mid J_{pi}\right) = \frac{\pi_{sk} P\left(J_{pi} \mid k\right)}{\sum_{k=1}^{C} \pi_{sk} P\left(J_{pi} \mid k\right)} \tag{11}$$

Then the sum of posterior probabilities is calculated

$$S_k = \sum_{l=1}^{N} P_{l,sk}\left(k \mid J_{pi}\right) \tag{12}$$

The M-Step then updates the mixture parameters as follows:

$$\pi_{sk}^{'} = \frac{S_k}{N} \tag{13}$$

$$\mu_{sk}^{'} = \frac{\sum_{l=1}^{N} J_{ki} P_{l,sk}\left(k \mid J_{pi}\right)}{s_k} \tag{14}$$

$$\Sigma_{sk}^{'} = \frac{\sum_{l=1}^{N} \left(J_{pi} - \mu_{sk}^{'}\right)^T P_{l,sk}\left(k \mid J_{pi}\right)}{S_k} \tag{15}$$

The E-Steps and M-Steps are iterated until convergence.

By using the EM algorithm, we obtain the distribution of the Gabor jets at each landmark point. Then we use the mean of each Gaussian component instead of the mean of all jets.

4. Search Using Gabor Wavelets

We can estimate the displacement between pairs of jets up to 8 pixels apart. By comparing the jets of each feature, we can obtain the best fitting jet at a new position. Here we use a coarse to fine approach to search for local points.

4.1. Jet Displacement Estimation

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

For two point P_1 and P_2, if we know the Gabor jets of P_1 and P_2, and if the displacement of the two points is less than 8 pixels, we can use a coarse to fine grained approach to obtain the displacement of the two points. The procedure is as follows: Assume we know the coordinates of P_1 and the jet value J_2, the goal is to estimate the coordinates of P_2:

1. Set the frequency of the lowest level.
2. Calculate the Gabor coefficient of the current frequency level at the position of P_1, to obtain the vector J_1.
3. Calculate $\vec{d} = \vec{d}(J_1, J_2)$ using equation (7).
4. Calculate the new position $P' = P + \vec{d}_i$.
5. Increase the frequency and go to step 2, until the highest frequency is reached.

4.2. Point Displacement Estimation

For the *ith* landmark point, the initial position is P_i and the Gabor jet is J_i. Following Section 3.2, we obtain the mean of each Gaussian component μ_{sk} of the point, $k = 1 \cdots c_s$, where c_s is the number of Gaussian components. The point displacement is calculated as follows:

1. For each mean in each Gaussian component, calculate the displacement between J_i and μ_{sk}.

$$\vec{d}_k = \vec{d}(J_i, \mu_{sk}) \tag{16}$$

2. Get the new position candidates.

$$P'_{ik} = P_i + \vec{d}_k \tag{17}$$

3. Calculate the Gabor jets for each new position J'_{ik}.
4. Use the phase-insensitive similarity function (5) to calculate the similarity $S_{ak}(J'_{ik}, \mu_{sk})$ between the new jets and the mean jets in the corresponding Gaussian component.
5. Select the highest similarity value from $S_{ak}(J'_{ik}, \mu_{sk})$ (a total of c_s similarity values). The new position is chosen as the corresponding new search point position.

By conducting this procedure, we move the original point to a new position which is most "similar" to the training model using the Gabor representation.

4.3. W-ASM Search Procedure

Our full search procedure is similar to the ASM method , except for the search for local points. The complete iterative procedure is as follows:

1. Use the face detection algorithm to detect the face and initialize the shape Y.
2. Generate the model instance $x = \overline{x} + P_s b$.
3. Use the method in Section 4.2 to search for each local point and obtain the new shape Y'.
4. Find the additional pose and shape parameter changes required to move x to the new search shape Y'.
5. Update the model parameters to match to Y'.
6. Apply the constraints on b.
7. Repeat step 2 until convergence.

5. Experimental Results

We manually labeled 515 pictures, each of size 200×200. On each image 87 landmarks are labeled. We select 400 images as the training set and the others as the test images. We compare the distance between each search shape and the manually labeled shapes.

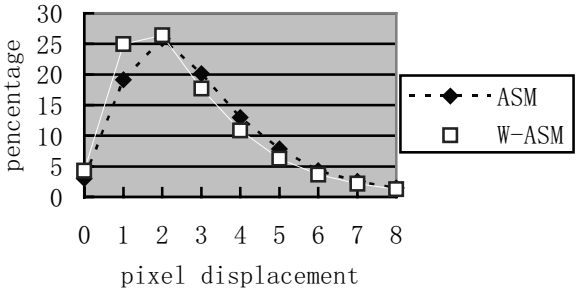

Figure 3. Point displacement test results.

First we calculate the displacement between each estimated point location and the corresponding labeled point to get the result shown in Figure 3. The x-coordinate is the average displacement (in pixels) between the estimated points and target point locations. The y-coordinate is the percentage of points whose displacement to the target is x. We can see that W-ASM achieves more accurate results than ASM.

For each test image, we calculate the overall displacement of the search shape to the labeled shape. The distance of two shapes is defined as follows:

$$Dis = \sum_{j=0}^{P} \sqrt{(x_{j1} - x_{j2})^2 + (y_{j1} - y_{j2})^2} \tag{18}$$

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

where P is the total number of landmarks (87 in our system). For each test image, we calculate *DisA* (the distance between ASM search shapes and the labeled shapes) and *DisW* (the distance of W-ASM search shapes to the labeled shapes). Then we calculate the value

$$m = (DisA - DisW)/DisA \times 100\% \qquad (19)$$

which measures the percentage of improvement of *DisW*. When $m<0$, that is $DisW>DisA$, this means that the result of W-ASM is worse than ASM. In Table 1 below, we can see that W-ASM works worse in 6 test images, and works better than ASM in the remaining 94 test images.

We also tested the algorithm on other face databases, including the CMU face database and the FERET database, which demonstrate significant variation in pose and illumination. Some of the search results are shown in Figures 5-12.

Our algorithm is tested on a P-III 450 computer with 256M memory. The average time to process a face image with *W-ASM* is about 0.5 to 0.8 seconds, while it takes about 0.2 to 0.4 seconds to process a face image using the ASM algorithm.

Table 1. Overall displacement comparison

m range (%)	the number of images
$m<-5$	2
$-5<m<0$	4
$0<m<5$	15
$5<m<10$	14
$10<m<15$	12
$15<m<20$	21
$20<m<25$	17
$25<m<30$	9
$30<m$	6

6. Conclusion

In this paper, we have presented the W-ASM algorithm, which uses Gabor features to model the local structure of face images. By using Gabor wavelets, the facial features can be more accurately located, compared to approaches using a simple gray value representation. An E-M algorithm is used to model the Gabor feature distribution, and a coarse–to-fine approach is used to search for local landmark features. Experimental results demonstrate the ability of W-ASM to accurately align and locate facial features.

The Elastic Bunch Graph Matching algorithm is time-consuming because it searches on the whole image, and uses translation, scale, aspect ratios and local distortions of the image grids to guide the search. This method requires significant computation. It is reported in [3] that when running on a SPARC station 10-512, it takes less than 30 seconds to extract features from one image. By contrast, the *W-ASM* algorithm, uses a statistical model to restrict shape variation and thereby limits the search space to a very small range. Thus the search procedure is much more efficient than ASM, requiring much less search time in practice (about 0.5-0.8 seconds for each image).

7. Reference

[1] M. Kass, A. Witkin, and D. Terzopoulos. Snakes, "Active contour models." *1st International Conference on Conference on Computer Vision*, London, June 1987, pp. 259-268.

[2] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, Jan 1990, pp. 103-108

[3] L. Wiskott, J.M. Fellous, N. Krüger, Cvd. Malsburg, "Face Recognition by Elastic Graph Matching," *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, eds. L.C. Jain et al., publ. CRC Press, ISBN 0-8493-2055-0, Chapter 11, 1999, pp. 355-396.

[4] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active Shape Models- their training and application." *Computer Vision and Image Understanding*, vol.61, no.1, Jan 1995, pp. 38-59.

[5] 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, 2001, pp. 681-685.

[6] J. G. Daugman, , "Complete discrete 2-D Gabor transform by neural network for image analysis and compression." *IEEE Trans. On Acoustics, Speech and Signal Processing*, vol 36, no. 7, July 1988, pp 1169-1179.

[7] D.U. Fleet, and A.D. Jepson, "Computation of component image velocity from local phase information." *International Journal of Computer Vision*, vol. 5, 1990, pp. 77-104.

[8] W.M. Theimer, and H.A Mallot, , "Phase-based binocular vergence control and depth reconstruction using active vision." *CVGIP*(60), No. 3, November 1994, pp. 343-358.

[9] R.N. Bracewell, *The Fourier Transform and Its Application*, New York, McGraw-Hill, 1978.

[10] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society*, Series B, vol. 39, no. 1, 1977, pp. 1--38.

[11] H. Akaike , "Information theory and an extension of the maximum likelihood principle." *Second International Symposium on Information Theory*, Budapest, 1973, pp. 267-281.

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

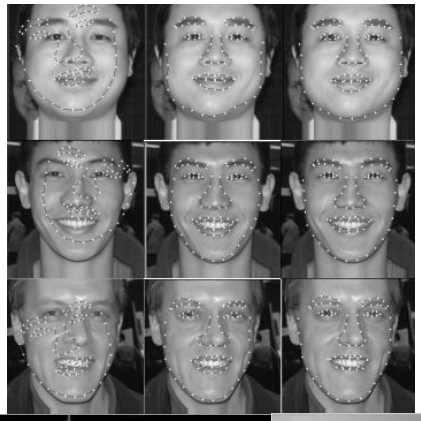

Figure 4. Comparison of the ASM and W-ASM algorithms. The first column contains the original images with initial shapes. The middle column contains the search results using ASM. The third column contains the search results using W-ASM.

Figure 5.

Figure 6.

Figure 7.

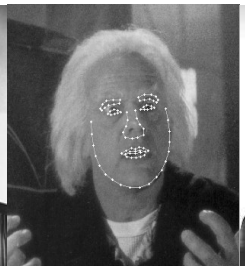

Figure 8.

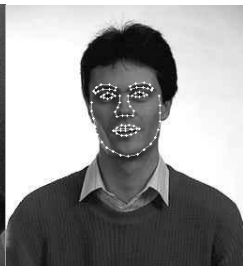

Figure 9.

Figure 10.

Figure 11.

Figure 5-11. Some search results using the W-ASM algorithm.

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE