ELSEVIER

Image and Vision Computing 21 (2003) 69–75

image AND vision COMPUTING

www.elsevier.com/locate/imavis

Face alignment using texture-constrained active shape models☆

Shuicheng Yan[a,*], Ce Liu[b], Stan Z. Li[b], Hongjiang Zhang[b], Heung-Yeung Shum[b], Qiansheng Cheng[a]

[a]*Department of Info. Sci., School of Math. Sci., Peking University, Peking 100871, People's Republic of China*
[b]*Microsoft Research Asia, Beijing Sigma Center, Beijing 100080, People's Republic of China*

Abstract

In this paper, we propose a texture-constrained active shape model (TC-ASM) to localize a face in an image. TC-ASM effectively incorporates not only the shape prior and local appearance around each landmark, but also the global texture constraint over the shape. Therefore, it performs stable to initialization, accurate in shape localization and robust to illumination variation, with low computational cost. Extensive experiments are provided to demonstrate our algorithm.
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Shape localization; Statistical shape model; Statistical texture model; Texture-constrained shape model

1. Introduction

Accurate extraction and alignment of faces from images are required in many computer vision and pattern recognition applications. Active Shape Models (ASM) and Active Appearance Models (AAM), proposed by Cootes et al. [4], are two popular shape and appearance models for object localization. They have been developed and improved for years [5–7,9].

In ASM, the local appearance model, which represents the local statistics around each landmark, efficiently finds the 'best' candidate point for each landmark in searching the image. The solution space is constrained by the properly trained global shape model. By means of modeling of the local features, ASM obtains nice results in shape localization. AAM [2,3,10] combines constraints on both shape and texture in its characterization of face appearance. In the context of this paper, *texture* means the intensity patch contained in the shape after warping to the mean shape [4]. 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 accurately in shape localization while AAM gives a better match to image texture. On the other hand, ASM tends to be stuck in local minima, dependent on the initialization. AAM is sensitive to the illumination, in particular if the lighting in the test is significantly different from the training. Meanwhile, training an AAM model is time consuming.

In this paper, a novel shape model, called Texture-Constrained Active Shape Model (TC-ASM), is proposed to address the above problems of ASM and AAM. TC-ASM inherits the local appearance model in ASM for the robustness of varying lighting. We borrow the global texture in AAM to TC-ASM, acting as a constraint over shape and providing an optimization criterion for determining the shape parameters. In TC-ASM, the conditional distribution of a shape given its associated texture is modeled as a Gaussian distribution. Thus, the texture corresponding to the shape obtained from the local appearance model, could linearly predict a *texture-constrained* shape. It converges to a local optimum when the shape from the local appearance model is very close to the texture-constrained shape.

Extensive experiments show that TC-ASM outperforms ASM and AAM in facial shape localization. It is also demonstrated that TC-ASM performs no worse than AAM in texture reconstruction.

This paper is organized as follows. In Section 2, we briefly review the shape and appearance models. The details of TC-ASM are discussed in Section 3. Experiments are presented in Section 4. We conclude this paper in Section 5.

☆ The work was performed at Microsoft Research Asia.

* Corresponding author.
E-mail address: scyan@msrchina.research.microsoft.com (S. Yan).

0262-8856/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0262-8856(02)00136-1

2. Classical shape and appearance models

Assume a training set of shape-texture pairs to be $\mathbf{\Omega} = \{(S_i, T_i^s)\}_{i=1}^N$. The shape $S_i = \{(x_j^i, y_j^i)\}_{j=1}^K \in \mathbb{R}2K$ is a sequence of K points in the image lattice. The texture T_i^s is the image patch enclosed by S_i. Let $\bar{S}$ be the mean shape of all the training shapes, as illustrated in Fig. 1. $\bar{S}$ is calculated from an iterative procedure [1] such that all the shapes are aligned to the tangent space of the mean shape $\bar{S}$. After shape warping [4], the texture T_i^s is warped correspondingly to $T_i \in \mathbb{R}^L$, , where L is the number of pixels enclosed by the mean shape $\bar{S}$. The warped textures are also aligned to the tangent space of the mean texture $\bar{T}$ in the same approach as computing $\bar{S}$.

2.1. ASM

In ASM, a shape is represented as a vector s in the low dimensional shape eigenspace $\mathbb{R}^k$, spanned by k ($< 2K$) principal modes (major eigenvectors) learned from the training shapes. A shape S could be linearly obtained from shape eigenspace:

$$S = \bar{S} + \mathbf{U}s, \tag{1}$$

where $\mathbf{U}$ is the matrix consisting of k principal modes of the covariance of $\{S_i\}$.

The local appearance models, which describe local image feature around each landmark, are modeled as the first derivatives of the sampled profiles perpendicular to the landmark contour [4]. For the jth landmark ($j = 1, \ldots, K$), we can derive the mean profile $\bar{g}_j$ and the covariance matrix $\mathbf{\Sigma}_j^g$ from the jth profile examples directly. At the current position $(x_j^{(n-1)}, y_j^{(n-1)})$ of the jth landmark, the local appearance models find the 'best' candidate (x_j^n, y_j^n) in the neighborhood $N(x_j^{(n-1)}, y_j^{(n-1)})$ surrounding $(x_j^{(n-1)}, y_j^{(n-1)})$, by minimizing the energy:

$$(x_j^n, y_j^n) = \underset{(x,y)\in N(x_j^{(n-1)}, y_j^{(n-1)})}{\arg\min} \|g_j(x,y) - \bar{g}_j\|^2_{\mathbf{\Sigma}_j^g} \tag{2}$$

where $g_j(x, y)$ is the profile of the jth landmark at (x, y) and $\|X\|^2_{\mathbf{\Lambda}} = X^T\mathbf{\Lambda}^{-1}X$ is the Mahalanobis distance measure with respect to a real symmetric matrix $\mathbf{\Lambda}$.

After relocating all the landmarks using the local appearance models, we obtain a new candidate shape S_{lm}^n. The solution in shape eigenspace is derived by maximizing the likelihood:

$$s^n = \arg\max_s p(S_{lm}^n|s) = \arg\min_s Eng(S_{lm}^n; s), \tag{3}$$

where[1]

$$Eng(S_{lm}^n; s) = \lambda\|S_{lm}^n - S'_{lm}n\|^2 + \|s_{lm}^n - s\|^2_{\mathbf{\Lambda}}. \tag{4}$$

[1] This is a deviation of the most used energy function with a squared Euclidean distance between S_{lm}^n and shape $S \in \mathbb{R}^{2K}$ derived from parameter s. It is more reasonable to take into account the prior distribution in the shape space.

In above equation, $s_{lm}^n = U^T(S_{lm}^n - \bar{S})$ is the projection of S_{lm}^n to the shape eigenspace, $S'_{lm}n = \bar{S} + Us_{lm}^n$ is the reconstructed shape, $\mathbf{\Lambda}$ is the diagonal matrix of the largest eigenvalues of the training data $\{S_i\}$. The first term is the squared Euclidean distance from S_{lm}^n to the shape eigenspace, and the second is the squared Mahalanobis distance between s_{lm}^n and s. λ balances the two terms.

Using the local appearance models leads to fast converge to the local image evidence. However, since they are modeled based on the local features, and the 'best' candidate point is only evaluated in local neighborhood, the solution of ASM is often suboptimal, dependent on the initialization.

2.2. AAM

In AAM, the texture eigenspace is spanned by the ℓ principle modes of $\{T_i\}$. The texture model is similar to the shape model:

$$T = \bar{T} + \mathbf{V}t \tag{5}$$

where $\mathbf{V}$ is the matrix consisting of ℓ principal orthogonal modes of the covariance in $\{T_i\}$, and t is the vector of texture parameters.

Let appearance $a = (\gamma s^T, t^T)^T$ [4] be a weighted vector of shape parameter s and texture t with the weighting parameter γ. AAM assumes that both the appearance displacement δa and the position (including the centroid (x, y), scale s and orientation θ) displacement δp are linearly dependent on the texture reconstruction error δT :

$$\delta a = \mathbf{A}_a\delta T, \qquad \delta p = \mathbf{A}_p\delta T \tag{6}$$

δT is continuously minimized in AAM by shifting the shape and position parameters as in Eq. (6). However, due to the high dimensions of the space T and a, the training of $\mathbf{A}_a$ and $\mathbf{A}_p$ is time and memory consuming. Meanwhile, since illuminations do not compose the image intensity linearly, δT could not accurately predict δa or δp under irregular lighting via linear mapping. Therefore, AAM solutions are often affected by varying illuminations.

3. Texture constrained active shape model

From above analysis, it is natural to develop a novel model to inherit the merits and reject the demerits of ASM and AAM. We propose a TC-ASM to borrow local appearance models from ASM for landmark localization, and incorporate the global texture constraint over the shape from AAM for more accurate shape parameters estimation. It consists of several types of models: a shape model, a texture model, K local appearance models, and a texture-constrained shape model. The former three types are exactly the same as in ASM and AAM. The texture-constrained shape model, or the mapping from texture to the expected shape, is simply assumed linear and could be easily learnt.

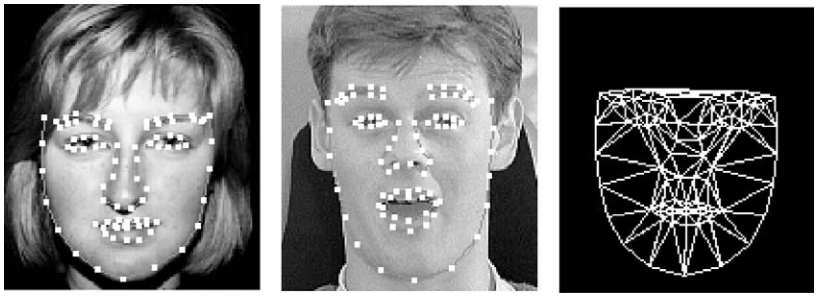

Fig. 1. Left and middle: two face instances labeled with 83 landmarks. Right: the mesh of the mean shape.

In each step of the optimization, a better shape is found under Bayesian framework. The details of the model will be introduced in the following.

3.1. Texture-constrained shape model

In the shape model, there are some landmarks defined on the edges or contours. Since they have no explicit definition for their positions, there exists uncertainty of the shape given the texture, whilst there are may be correlations between the shape and the texture. To formulate the correlations, the conditional distribution of shape parameters s given texture parameters t is simply assumed Gaussian, i.e.

$$p(s|t) \sim N(s_t, \mathbf{\Sigma}_t), \tag{7}$$

where $\mathbf{\Sigma}_t$ stands for the covariance matrix of the distribution, and s_t is linearly determined by the texture t. The linear mapping from t to s_t is:

$$s_t = \mathbf{R}t, \tag{8}$$

where $\mathbf{R}$ is a projection matrix that can be pre-computed from the training pairs $\{(s_i, t_i)\}$ by singular-value decomposition. For simplicity, $\mathbf{\Sigma}_t$ is assumed to be a known constant

Fig. 2. The comparison of the manually labeled shape (middle row) and the shape (bottom row) derived from the enclosed texture using the learned projection matrix: $s_t = \mathbf{R}t$. In the top row are the original images. All the images are test data.

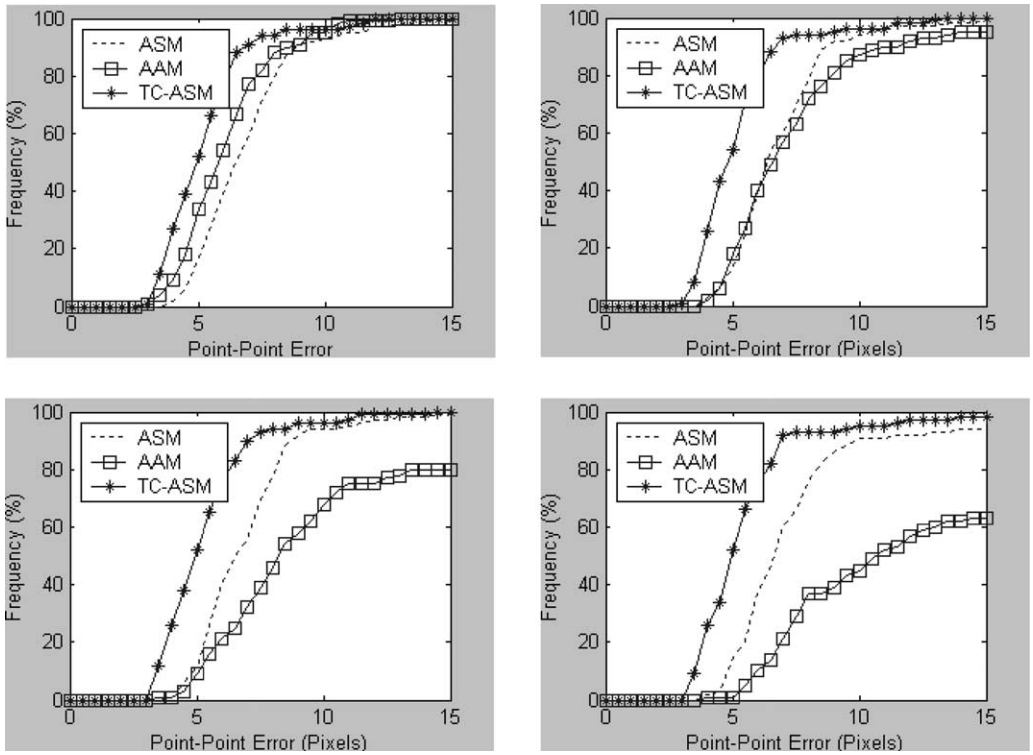

Fig. 3. Accuracy of ASM, AAM, TC-ASM. From upper to lower, left to right are the results obtained with the initial displacements of 10, 20, 30 and 40 pixels.

matrix. Fig. 2 demonstrates the accuracy of the prediction in the test data via the matrix **R**. We may see that the predicted shape is close to the labeled shape even under varying illuminations. Thus, the constraints over the shape from the texture can be used as an evaluation criterion in the shape localization task. The prediction of matrix **R** is also affected by illumination variation, yet since Eq. (8) is formulated based on the eigenspace, the influence of the unfamiliar illumination can be alleviated when the texture is projected to the eigenspace.

The distribution (Eq. (7)) can also be represented as the prior distribution of s given the shape s_t :

$$p(s|s_t) \propto \exp\{ -Eng(s; s_t)\}, \tag{9}$$

where the energy function is:

$$Eng(s; s_t) = \|s - s_t\|^2_{\Sigma_t}. \tag{10}$$

3.2. TC-ASM in Bayesian framework

TC-ASM search starts with the mean shape, namely the shape parameters $s^0 = 0$. The whole search process is outlined as below:

(1) Set the iteration number $n = 1$;
(2) Using the local appearance models in ASM, we may

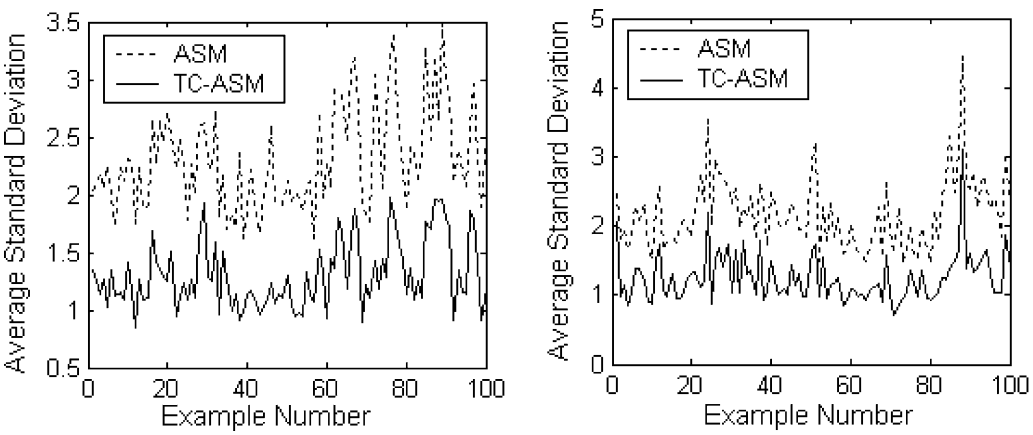

Fig. 4. Standard deviation in the results of each example for ASM (dotted) and TC-ASM (solid) with training set (left) and test set (right).

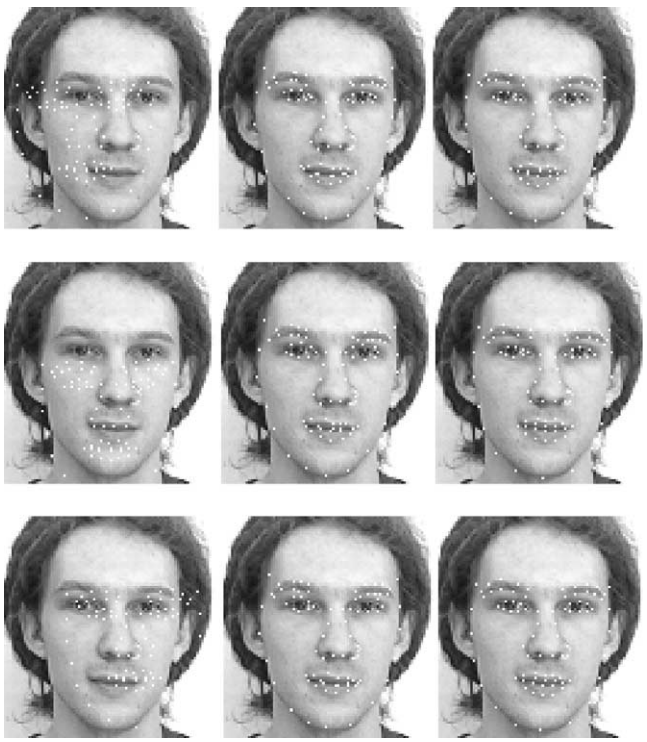

Fig. 5. Stability of ASM (middle column) and TC-ASM (right column) in shape localization. The different initialization conditions are showed in the left column. Note that there are less variations among the positions of the eyebrows and mouth points in TC-ASM.

obtain the candidate shape S_{lm}^n with the shape parameters s_{lm}^n based on the shape $S^{(n-1)}$ of the previous iteration;

(3) The texture enclosed by S_{lm}^n is warped to the mean shape, denoted by t^n. The texture-constrained shape s_t^n is predicted from t^n by Eq. (8);

(4) The *posterior* (MAP) estimation of S^n or s^n given S_{lm}^n and s_t^n is derived based on the Bayesian framework;

(5) If the stopping condition is satisfied, exit; otherwise, let $n = n + 1$, goto step 2.

In the following, we illustrate the step 4 and the stopping condition in detail. To simplify the notation, we shall omit the superscript n in following deduction since the iteration number is constant. In step 4, the posterior (MAP) estimation of s given S_{lm} and s_t is:

$$p(s|S_{lm}, s_t) = \frac{p(S_{lm}|s, s_t)p(s, s_t)}{p(S_{lm}, s_t)}. \tag{11}$$

Assume that S_{lm} is conditionally independent to s_t, given s, i.e.

$$p(S_{lm}|s, s_t) = p(S_{lm}|s). \tag{12}$$

Then

$$p(s|S_{lm}, s_t) \propto p(S_{lm}|s)p(s|s_t). \tag{13}$$

The corresponding energy function is:

$$Eng(s; S_{lm}, s_t) = Eng(S_{lm}; s) + Eng(s; s_t) \tag{14}$$

From Eqs. (4) and (10), the best shape obtained in each step is

$$\begin{aligned} s &= \arg \min_s [Eng(s; S_{lm}) \Rightarrow Eng(S_{lm}; s)] \\ &= \arg \min_s \|s_{lm} - s\|_{\mathbf{\Lambda}}^2 + \|s - s_t\|_{\mathbf{\Sigma}_t}^2 \\ &= \arg \min_s [s^T(\mathbf{\Lambda}^{-1} + \mathbf{\Sigma}_t^{-1})s - 2s^T(\mathbf{\Lambda}^{-1}s_{lm} + \mathbf{\Sigma}_t^{-1}s_t)] \\ &= (\mathbf{\Lambda}^{-1} + \mathbf{\Sigma}_t^{-1})^{-1}(\mathbf{\Lambda}^{-1}s_{lm} + \mathbf{\Sigma}_t^{-1}s_t). \end{aligned} \tag{15}$$

After restoring the superscript of iteration number, the best shape obtained in step n is

$$s^n = (\mathbf{\Lambda}^{-1} + \mathbf{\Sigma}_t^{-1})^{-1}(\mathbf{\Lambda}^{-1}s_{lm}^n + \mathbf{\Sigma}_t^{-1}s_t^n). \tag{16}$$

This indicates that the best shape derived in each step is a weighted average between the shape from the local appearance model and the texture-constrained shape. In this sense, TC-ASM could be regarded as a trade-off between ASM and AAM methods.

The stopping condition of the optimization is: if the shape from the local appearance model and the texture-constrained shape are the same, i.e. the solution generated by ASM is verified in AAM, the optimal solution must have been touched. In practice, however, these two shapes would

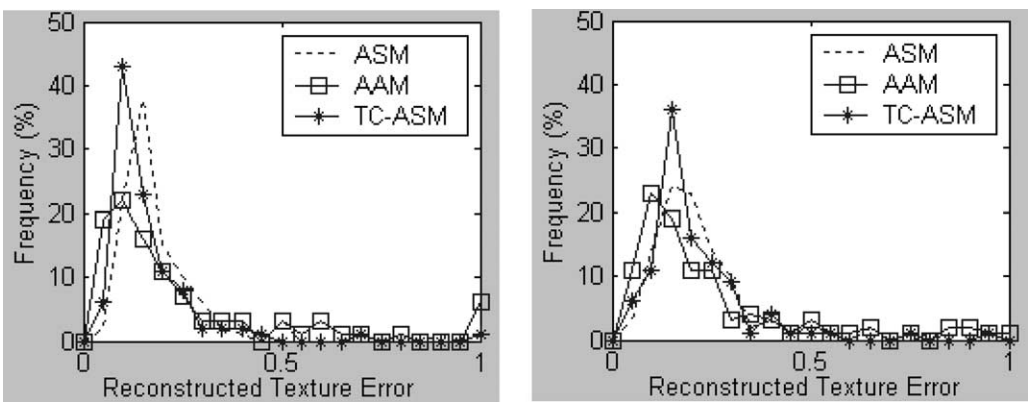

Fig. 6. Distribution of the texture reconstruction error in ASM (dotted), AAM (square) and TC-ASM (asterisk) with training data (left) and test data (right).

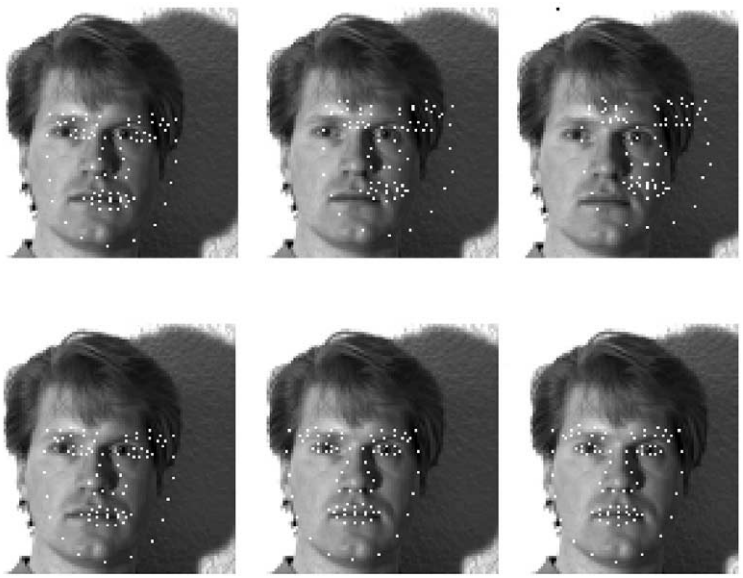

Fig. 7. Sensitivities of AAM (upper) and TC-ASM (lower) to illumination condition not seen in the training data. From left to right are the results obtained at the 0th, 2nd, and 10th iterations. Note that the result in different level of image pyramid is scaled back to the original scale.

hardly turn to be the same. A threshold is introduced to evaluate the similarity and sometimes the convergence criterion in ASM is used (if the above criterion has not been satisfied for a long time). For higher efficiency and accuracy, a multi-resolution pyramid method is adopted in optimization process.

4. Experiments

A data set containing 700 face images from about 300 persons with different illumination conditions and expressions are selected from the AR database [8] in our experiments, each of which is a 512×512, 256 gray-levels image containing the frontal view face about 200×200. Eighty-three landmark points are manually labeled on the face. We randomly select 600 for training and the other 100 for testing.

For comparison, ASM and AAM are trained on the same data sets, in a three-level image pyramid (Resolution is reduced 1/2 level by level) as TC-ASM. By means of PCA with 98% total variations retained, the dimension of the shape parameter in ASM shape space is reduced to 88, and the texture parameter vector in AAM texture space is reduced to 393. The concatenated vector of the shape and texture parameter vector with the weighting parameter [2]

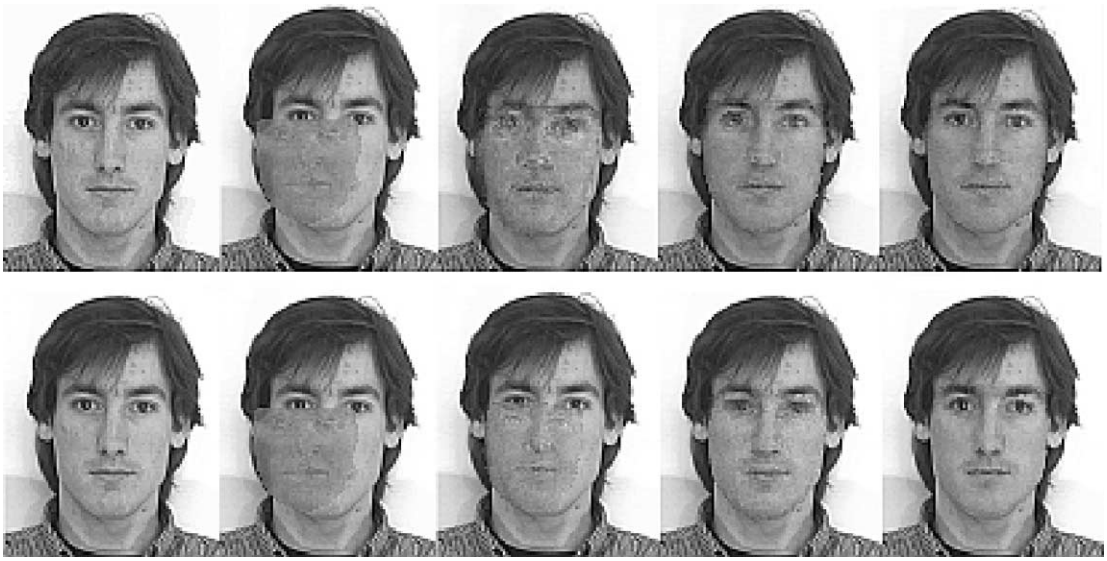

Fig. 8. Scenarios of AAM (upper) and TC-ASM (lower) alignment with texture reconstruction error 0.3405 and 0.1827, respectively. From left to right are the original image and the results obtained at the 0th, 5th, 10th, 15th iterations. Note that the result in different level of image pyramid is scaled back to the original scale.

$\gamma = 13.77$ is reduced to 277. Two types of experiments are presented: (1) the comparison of the point-position accuracy and (2) the comparison of the texture reconstruction error.

4.1. Point position accuracy

The average point–point distances between the searched shape and the manually labeled shape of the three models are compared in Fig. 3. The vertical axis represents the percentage of the solutions for which the average point–point distances to the manually labeled ones are smaller than the corresponding horizontal axis value. The statistics are calculated from 100 test images with different initializations, with random displacements to the ground truth of 10, 20, 30 and 40 pixels. The results show that TC-ASM outperforms both ASM and AAM in most cases since the curve of TC-ASM lies above the curves for ASM and AAM. It also suggests that AAM outperforms ASM when the initial displacement is small, while ASM is more robust to the increasing of the initial displacement.

We compare the stability, which is measured by the standard deviation of the results from the initializations with similar point–point distances between the initial shapes and the ground truth, of TC-ASM with ASM in Fig. 4. The value of horizontal axis is the index number of the selected examples, whereas the value of the vertical axis is the average standard deviation of the results obtained from 10 different initializations which deviate from the ground truth by approximately 20 pixels. The result convinces that TC-ASM is more stable to initializations. An example is given in Fig. 5.

4.2. Texture reconstruction error

The texture reconstruction error comparison of the three models in Fig. 6 illustrates that TC-ASM improves the accuracy of the texture matching. The texture accuracy of TC-ASM is close to that of AAM while its position accuracy is better than AAM (see Fig. 3). Although AAM has more cases with small texture reconstruction error, TC-ASM has more cases with the texture reconstruction error smaller than 0.2.

An example in which AAM fails for a different illumination condition from the training data, yet TC-ASM performs well is presented in Fig. 7 . Fig. 8 shows a scenario of AAM and TC-ASM alignment.

From the experiment, TC-ASM is more computationally expensive than ASM, but it is much faster than AAM. In our experiment (600 training images, 83 landmarks and a P-III 667 computer with 256 M memory), it takes averagely 32 ms per iteration, which is twice of ASM (16 ms) but one fifth of AAM (172 ms). It takes TC-ASM about three iterations per level to converge.

5. Conclusion

In this paper we proposed a novel shape model, TC-ASM, for face shape localization. TC-ASM efficiently incorporates the local information around each landmark and the global texture information for alignment. It is more robust to initialization, more accurate in shape localization and less sensitive to illumination, when compared with conventional methods. For future work, the generalization of the shape prediction from the texture can be evaluated on a larger data set.

References

[1] T. Cootes, G. Edwards, C. Taylor, Active appearance models, Proceedings of the European Conference on Computer Vision 2 (1998) 484–498.
[2] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 681–685.
[3] T. Cootes, C. Taylor, On representing edge structure for model matching, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1 (2001) 1114–1119.
[4] T. Cootes, C. Taylor, Statistical models of appearance for computer vision, http://www.isbe.man.ac.uk/~bim/refs.html, 2001.
[5] B. Ginneken, A. Frangi, J. Staal, B. Romeny, M. Viergever, A non-linear gray-level appearance model improves active shape model segmentation, Proceedings of the Mathematical Methods in Biomedical Image Analysis, 2001.
[6] X.W. How, S.Z. Li, H.J. Zhang, Direct appearance models, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Hawaii December (2001) 828–833.
[7] Y. Li, S. Gong, H. Liddell, Modelling faces dynamically across views and over time, Proceedings of the IEEE International Conference on Computer Vision, Canada July (2001) 554–559.
[8] A. Martinez, R. Benavente, The AR face database, Technical Report 24, CVC, June 1998.
[9] S. Mitchell, B. Lelieveldt, R. Geest, J. Schaap, J. Reiber, M. Sonka, Segmentation of cardiac MR images: an active appearance model approach, SPIE Medical Imaging February (2000).
[10] M. Stegmann, Active appearance models: theory, extensions and cases, Master's thesis, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark, 2000.