Sparse Shape Registration for Occluded Facial Feature Localization

Fei Yang, Junzhou Huang and Dimitris Metaxas

***Abstract*— This paper proposes a sparsity driven shape registration method for occluded facial feature localization. Most current shape registration methods search landmark locations which comply both shape model and local image appearances. However, if the shape is partially occluded, the above goal is inappropriate and often leads to distorted shape results. In this paper, we introduce an error term to rectify the locations of the occluded landmarks. Under the assumption that occlusion takes a small proportion of the shape, we propose a sparse optimization algorithm that iteratively approaches the optimal shape. The experiments in our synthesized face occlusion database prove the advantage of our method.**

I. INTRODUCTION

Automatic face registration plays an important role in many face identification and expression analysis algorithms. It is a challenging problem for real world images, because various face shapes, expressions, poses and lighting conditions greatly increase the complexity of the problem.

Many current shape registration algorithms are based on statistical point distribution models. A shape is described by 2D or 3D coordinates of a set of labeled landmarks. These landmarks are predefined as points located on the outline, or some specific positions (e.g., eyes pupils). These algorithms work by modeling how the labeled landmarks tend to move together as the shape varies. Cootes et al. [3][6] first presented Active Shape Models (ASM) using linear shape subspaces. This method assumes that the residuals between model fit and images have a Gaussian distribution. There have been many modifications to the classical ASM. Cootes et al. [5] built shape models using a mixture of Gaussian. Romdhani et al. [16] used Kernel PCA to generate nonlinear subspaces. Other improvements including Rogers and Graham [15], Van Ginneken et al. [8][11], Jiao et al. [10], Li and Ito [12]. Milborrow et al. [14] etc. Cootes et al. [4] also proposed Active Appearance Models, which merges the shape and profile model of the ASM into a single model of appearance, and itself has many descendants.

If parts of the shape are occluded, the unobservable landmarks cannot find a correct match. The previous methods based on ASM can not handle this problem because the incorrect matches are projected into the shape space, which often leads to distorted shape results. Some other shape models tried to alleviate this problem. Zhou et al. [18] proposed a Bayesian inference solution based on tangent shape approximation. Gu and Kanade [9] used a generative model and EM-based algorithm to implement the maximum a posterior. Felzenszwalb et al. [7] and Tan et al. [17] applied pictorial structures which model the spacial relationship between parts of objects. However, shape registration under occlusions has not been directly modeled and is far from being resolved.

Fig. 1. Faces with occlusion

In this paper, we propose a new shape registration method to directly handle this problem. We extend the linear subspace shape model by introducing an error term to rectify the locations of the occluded landmarks. With the assumption that occlusion takes a small proportion of the shape, the error term is constrained to be sparse. The proposed method iteratively approximates the optimal shape. To quantitatively evaluate the proposed method, we built three face datasets with synthesized occlusions. Our experimental results prove the advantage of our method.

The rest of this paper is organized as the follows. Section 2 presents the mathematical formulations and proposes our algorithm. Section 3 illustrates experimental results. Section 4 concludes.

II. SPARSE SHAPE REGISTRATION

Given a shape containing N landmarks, the shape vector S is defined by concatenating x and y coordinates of all the landmarks.

$$S = [x_1, y_1, x_2, y_2, ..., x_N, y_N]^T \quad (1)$$

This work was supported by the National Space Biomedical Research Institute through NASA NCC 9-58

Fei Yang and is with the Department of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ, 08854, USA, feiyang@cs.rutgers.edu

Junzhou Huang is with the Department of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ, 08854, USA, jzhuang@cs.rutgers.edu

Dimitris Metaxas is with Faculty of the Department of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ, 08854, USA, dnm@cs.rutgers.edu

We assume the shape is a linear combination of m shape basis

$$S = \bar{S} + b_1 u_1 + b_1 u_1 + \cdots + b_m u_m \quad (2)$$
$$= \bar{S} + Ub \quad (3)$$

where U is a matrix with size n by m containing m shape basis. b is a m by 1 vector for the coefficients. A shape registration method seeks to locate landmarks complying both the shape model and image features. If some landmarks are occluded, the correct postion will not get a high responses from the appearance templates. It means that, the high response positions that best matching templates are not the real positions of these landmarks. Therefore, these incorrect positions should not be used for global shape matching.

We define an error term S_e to directly model the occluded landmark positions. The hidden shape vector S is the sum of the shape estimate $\hat{S}$ and shape error S_e.

$$S = \hat{S} + S_e \quad (4)$$

The shape transformation parameters (scaling, rotation and translation) are denoted by θ. The posterior likelihood of θ, shape parameter b, hidden shape vector S, error S_e given image I is

$$p(\theta, b, S, S_e|I) \propto p(\theta)p(b)p(S|b)p(S_e)p(I|\theta, S, Se) \quad (5)$$

the prior $p(\theta)$ can be considered as a constant, since there is no preference for shape scale, orientation and location. We take the negative logarithm of Equation (5). Now we aim to minimize the following energy function.

$$E = -\log p(b) - \log p(S|b) - \log p(S_e) - \log p(I|\theta, S, Se) \quad (6)$$
$$= E_b + E_S + E_{S_e} + E_I \quad (7)$$

We expand Equation (7).

$$E_b = \frac{1}{2} b^T \Lambda^{-1} b \quad (8)$$

where Λ is the m by m diagonal matrix containing the largest m eigenvalues of Σ. For simplicity, we consider the shape model to be a single Gaussian distribution with mean $\bar{S}$ and covariance Σ^{-1}. The shape basis U and Λ are computed from SVD decomposition to the covariance matrix $\Sigma = U\Lambda U^T$. We keep only the m eigenvectors as shape basis and largest m eigenvalues in Λ. The single Gaussian model can also be extended to a mixture of Gaussians following methods from Gu et al.[9].

The shape energy E_S can be written as

$$E_S = \frac{1}{2}||S - Ub - \bar{S}||^2 \quad (9)$$
$$= \frac{1}{2}||\hat{S} + S_e - Ub - \bar{S}||^2 \quad (10)$$

We assume that the occluded landmarks takes only a small proportion of all the landmarks, which means S_e is sparse. We define the energy term E_{S_e} as the L_1 norm of S_e, with a diagonal weighting matrix W.

$$E_{S_e} = \lambda \cdot ||WS_e||_1 \quad (11)$$

The image likelihood at each landmark position is assumed to be independent to each other. So that

$$p(I|\theta, S, Se) = \prod_{i=1}^{N} p(I_i|\theta, S, S_e) \quad (12)$$

We also use a single Gaussian model for the appearance at each landmark position. Thus the energy term E_I can be written as

$$E_I = \frac{1}{2}\sum_{i=1}^{N}(F(\mathbf{x}_i) - u_i)^T \Sigma_i^{-1}(F(\mathbf{x}_i) - u_i) \quad (13)$$
$$= \frac{1}{2}\sum_{i=1}^{N} d(\mathbf{x}_i)^2 \quad (14)$$

where $F(\mathbf{x}_i)$ is the feature extracted at landmark position $\mathbf{x}_i$ from shape $\hat{S}$; u_i and Σ_i are the mean and covariance of the Gaussian appearance model for landmark i. The energy term can be simply written as a sum of Mahalanobis distances $d(\mathbf{x}_i)$.

A. Iterative Optimization

Now we aim to minimize the energy function E

$$E = E_b + E_S + E_{S_e} + E_I \quad (15)$$

Firstly, we define E_p as the sum of E_b, E_S and E_{S_e}

$$E_p(b, S_e) = \frac{1}{2} b^T \Lambda^{-1} b + \frac{1}{2}||\hat{S} + S_e - Ub - \bar{S}||^2 + \lambda \cdot ||WS_e||_1 \quad (16)$$

E_p is a convex function, which can be minimized by gradient descent method. The first and second order partial derivatives of E_p to b and S_e are

$$\frac{\partial(E_b + E_S)}{\partial b} = \Lambda^{-1} b - U^T(\hat{S} + S_e - Ub - \bar{S}) \quad (17)$$
$$\frac{\partial(E_b + E_S)}{\partial S_e} = \hat{S} + S_e - Ub - \bar{S} \quad (18)$$
$$\frac{\partial^2 E_p}{\partial b^2} = (\Lambda^{-1} + I) \quad (19)$$
$$\frac{\partial^2 E_p}{\partial S_e^2} = I \quad (20)$$

The algorithm to minimize E_p is shown in Algorithm 1.

Algorithm 1 *Minimize* $Ep = E_b + E_S + E_{S_e}$

1: $b^0 = U^T(\hat{S} - \bar{S})$, $S_e^0 = 0$
2: **for** $k = 0 : k_{max}$ **do**
3: Compute L to be the largest eigenvalue of $\frac{\partial^2 E_p}{\partial b^2}$.
4: $b^{k+1} = b^k - \frac{1}{L} \cdot \frac{\partial E_p}{\partial b}$
5: $S_e^{k+\frac{1}{2}} = S_e^k - \frac{\partial E_p}{\partial S_e}$
6: $S_e^{k+1} = max(|S_e^{k+\frac{1}{2}}| - \lambda, 0) \cdot sign(S_e^{k+\frac{1}{2}})$
7: **end for**

Secondly, we try to minimize E_I. Notice that E_I is a discontinuous function. Traditional active shape model based

algorithms measure the image likelihood around the the landmarks, and move the landmark to the new position which has maximum response. For real world images, this method is sensitive to noises. Instead of using the single maximum response point, we use the kernel density estimation and mean shift method to find the position best matching the landmark.

Image gradient features are extracted at a set of n points $\{\mathbf{x}_{i,j}\}_{j=1...n}$ around a landmark at point $\mathbf{x}_i$. we define $f(\mathbf{x}_{i,j})$ as the square of Mahalanobis distance at point $\mathbf{x}_{i,j}$.

$$f(\mathbf{x}_{i,j}) = d(\mathbf{x}_{i,j})^2 \tag{21}$$

The kernel density estimation computed in the point $\mathbf{x}$, with kernel K and bank-width h, is given by

$$\hat{f}_{h,K}(\mathbf{x}) = \frac{1}{C}\sum_{j=1}^{n} f(\mathbf{x}_{i,j}) \cdot K(\frac{\mathbf{x}-\mathbf{x}_{i,j}}{h}) \tag{22}$$

Let G be profile of kernel K. When K is the normal kernel, its profile G has the same expression. As shown in [1], the gradient estimate at point $\mathbf{x}$ is proportional to the density estimate in $\mathbf{x}$ computed with kernel G and the mean shift vector computer with kernel G.

$$\hat{\nabla} f_{h,K}(\mathbf{x}) = C \cdot \hat{f_{h,G}}(\mathbf{x}) \cdot m_{h,G}(\mathbf{x}) \tag{23}$$

The mean shift vector $m_{h,G}(\mathbf{x})$ is defined as

$$m_{h,G}(\mathbf{x}) = \frac{\sum_{j=1}^{n} \mathbf{x}_{i,j} \cdot f(\mathbf{x}_{i,j}) \cdot G\left(||\frac{\mathbf{x}-\mathbf{x}_{i,j}}{h}||^2\right)}{\sum_{j=1}^{n} f(\mathbf{x}_{i,j}) \cdot G\left(||\frac{\mathbf{x}-\mathbf{x}_{i,j}}{h}||^2\right)} - \mathbf{x} \tag{24}$$

The local minimum of E_I can be acquired using gradient descent. We take steps proportional to the negative of the gradient, as shown in Algorithm 2.

Algorithm 2 *Minimize* E_I

1: **for** $i = 1 : N$ **do**
2: **for** $k = 0 : k_{max}$ **do**
3: Compute $\hat{\nabla} f_{h,K}(\mathbf{x}_i^k)$ using equation (23)
4: $\mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \hat{\nabla} f_{h,K}(\mathbf{x}_i^k)$
5: **end for**
6: **end for**

To minimize E, we alternately run Algorithm 1 and Algorithm 2. Our algorithm is shown in Algorithm 3.

III. Experiment

To evaluate our algorithm, we create a synthesized face occlusion database using face images from AR [13] database. The AR database contains frontal face images from 126 people. Each person has 26 images with different expressions, occlusions and lightening conditions. We select 509 face images from section 1,2,3,5 and use the 22 landmark positions provided by T.F.Cootes [2] as the ground truth. The landmark positions are shown in Fig. 2.

The occlusion masks are designed to simulate the occlusions most frequently seen in real world. As shown in Fig. 3. We design three types of masks: A cap mask is put above the eyes but occludes all eye brow regions; A hand mask is put on mouth which also occludes nose tip and part of chin; And a scarf mask is applied to occlude the mouth and chin. These masks are carefully located at the same position for all faces. By putting masks on clear face images, we still know the ground truth positions of all occluded landmarks, which is convenient for quantitative evaluation.

Algorithm 3 *Sparse Shape Optimization*

1: Compute θ using detection result
2: Initial status $b_0 = 0$, $S_e = 0$, $S = \bar{S}$, $\hat{S} = \bar{S}$, $\hat{S}' = M_\theta(\hat{S})$
3: **repeat**
4: Run Algorithm 2 to optimize $\hat{S}'$
5: Compute transformation parameter θ matching $\hat{S}'$ to $\bar{S}$
6: $\hat{S} = M_\theta^{-1}(\hat{S}')$
7: Run Algorithm 1 to optimize b and S_e
8: $\hat{S}' = M_\theta(\bar{S} + Ub)$
9: **until** $\hat{S}'$ converges

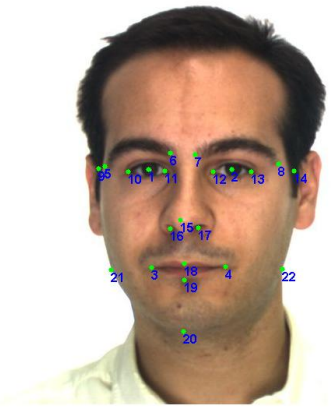

Fig. 2. Face image with 22 landmarks

The shape registration result for one testing image is shown in Fig. 4. The ground truth positions are marked using red stars. The result of ASM is shown in blue lines and the result of our method is shown in green lines. On the right side is the sparse shape error recovered during one iteration. The sparse coefficients on the left side are corresponding to the landmarks at contour of chin. And the ones on the right side are corresponding to the landmarks in mouth. In this figure, the landmark indexes are different from Fig. 2, because we use a linear shape model containing more landmarks than the ground truth.

In order to assess the localization precision, we apply the normalized error metric similar to Jesorsky et al. [**?**]. The normalized error for each point is defined as the Euclidean distance from the ground truth, normalized by the distance between eye centers. This metric is scale invariant.

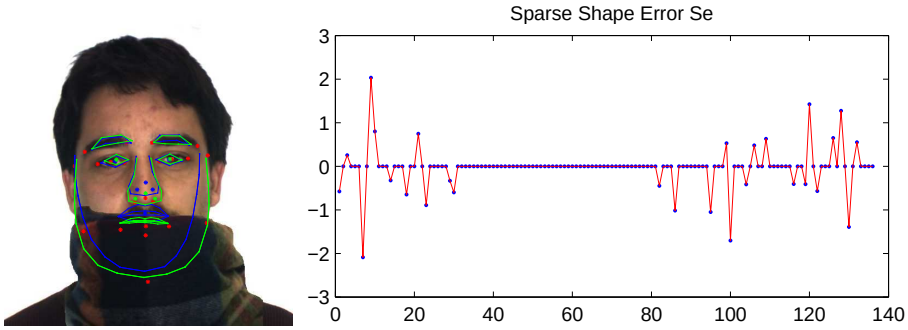

Fig. 4. Sparse shape error

Fig. 3. Faces with artificial occlusion

We compare our algorithm with the Milborrow's extended Active Shape Model [14], which was reported better performance than traditional ASM methods. The results are shown in Fig. 5. On the hat occlusion dataset, our method has significantly better localization accuracy for landmarks 6,7,8,9, which are the four landmarks at the ends of eye brows. On the hand occlusion dataset, our method has much better accuracy for landmarks 3,4,18,19 which are occluded landmarks on mouth, and landmarks 15, 16, 17 which are occlude landmarks on nose. On the scarf occlusion dataset, our method gets much better accuracy for landmarks 3,4,18,19 which are occluded landmarks on mouth, and landmarks 20,21,22 which are occluded landmarks on chins.

In all the three datasets, we decrease the normalized error of the occluded landmarks to level of 0.2, which is close to the error level of the non-occluded landmarks. The speed of our method is about 20 percent slower than Milborrow's extended ASM, because of extra cost to compute gradients. We set the maximum number of iterations k_{max} to be 2 in Algorithm 1 and 2. Our experiments with larger k_{max} do not

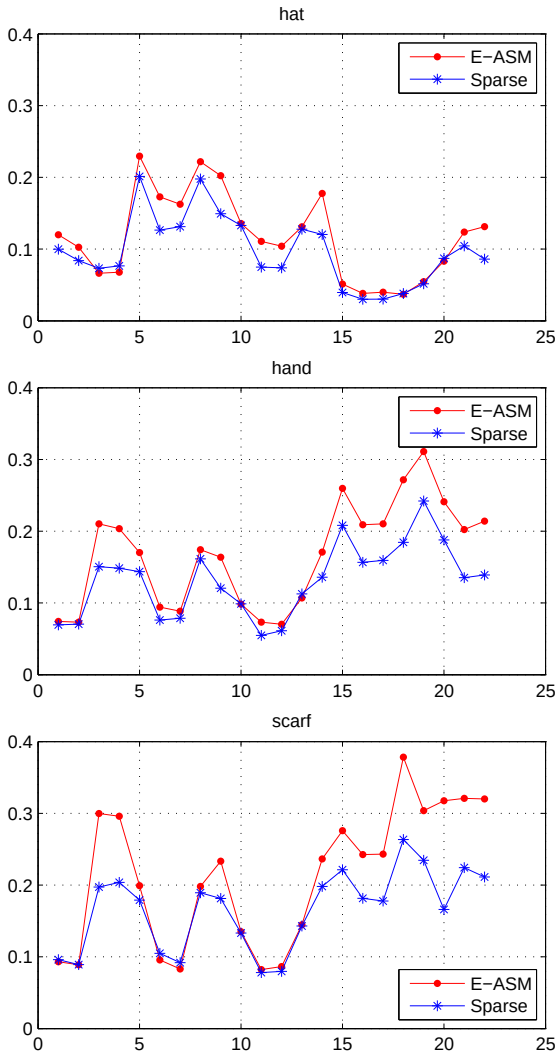

Fig. 5. Results on AR database

have significant better accuracy. We show more localization results in Fig. 6. The ground truth positions are marked as red stars. The ASM results are shown in blue lines and our results are shown in green lines.

IV. Conclusion

In this paper, we propose a sparsity driven shape registration method for occluded facial feature localization. By introducing a sparse error term into the linear shape model, our algorithm is more robust for feature localization, especially for the occluded landmarks. Extensive experiments in our synthesized face occlusion database prove the advantage of our method. Our future work includes creating more occlusion testing scenarios, and extend our algorithm to mixture shape models.

References

[1] D. Comaniciu, P. Meer, and S. Member. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24:603–619, 2002.

[2] T. F. Cootes. `http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/tarfd_markup/tarfd_markup.html`.

[3] T. F. Cootes, D. Cooper, C. J. Taylor, and J. Graham. A trainable method of parametric shape description. *Proc. British Machine Vision Conference*, pages 54–61, 1991.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proc. the 5th European Conference on Computer Vision (ECCV)*, pages 484–498, 1998.

[5] T. F. Cootes and C. J. Taylor. A mixture model for representing shape variation. In *Image and Vision Computing*, pages 110–119. BMVA Press, 1997.

[6] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Compututer Vision and Image Understanding*, 61(1):38–59, 1995.

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61, 2003.

[8] B. V. Ginneken, A. F. Frangi, R. F. Frangi, J. J. Staal, B. M. T. H. Romeny, and M. A. Viergever. Active shape model segmentation with optimal features. *IEEE Trans. Medical Imaging*, 21:924–933, 2002.

[9] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *Proc. the 10th European Conference on Computer Vision (ECCV)*, pages I: 413–426, 2008.

[10] F. Jiao, S. Li, H. Shum, and D. Schuurmans. Face alignment using statistical models and wavelet features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I: 321–327, 2003.

[11] J. J. Koenderink and A. J. van Doorn. The structure of locally orderless images. *International Journal of Computer Vision*, 31(2-3):159–168, 1999.

[12] Y. Li and W. Ito. Shape parameter optimization for adaboosted active shape model. In *Proc. the 10th IEEE International Conference on Computer Vision (ICCV)*, pages 251–258, 2005.

[13] A. Martinez and R. Benavente. The ar face database. Technical report.

[14] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *Proc. the 10th European Conference on Computer Vision (ECCV)*, pages 504–513, 2008.

[15] M. Rogers and J. Graham. Robust active shape model search for medical image analysis. In *Proc. International Conference on Medical Image Understanding and Analysis*, 2002.

[16] S. Romdhani, S. Gong, A. Psarrou, and R. Psarrou. A multi-view nonlinear active shape model using kernel pca. In *Proc. British Machine Vision Conference*, pages 483–492. BMVA Press, 1999.

[17] X. Tan, F. Song, Z. Zhou, and S. Chen. Enhanced pictorial structures for precise eye localization under incontrolled conditions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1621–1628, 2009.

[18] Y. Zhou, L. Gu, and H. Zhang. Bayesian tangent shape model: Estimating shape and pose parameters via bayesian inference. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I: 109–116, 2003.

Fig. 6. Some of the localization results