Face Alignment Using Local Hough Voting

Xin Jin Xiaoyang Tan Liang Zhou

Abstract— **We present a novel Hough voting-based method to improve the efficiency and accuracy of fiducial points localization, which can be conveniently integrated with any global prior model for final face alignment. Specifically, two or more stable facial components (e.g., eyes) are first localized and fixed as anchor points, based on which a separate local voting map is constructed for each fiducial point using kernel density estimation. The voting map allows us to effectively constrain the search region of fiducial points by exploiting the local spatial constraints imposed by it. In addition, a multi-output ridge regression method is adopted to align the voting map and the response map of local detectors to the ground truth map, and the learned transformations are then exploited to further increases the robustness of the algorithm against various appearance variations. Encouraging experimental results are given on several publicly available face databases.**

I. INTRODUCTION

In many face-related applications, such as face recognition, gaze detection and facial expression analysis, it is crucial to find out the semantic correspondence between facial features. This task, usually called face alignment, is popularly tackled by fitting a deformable template model with the given face image and then use the predefined arrangements of feature points in the model to obtain the needed correspondence in the image space. In this procedure, how to accurately and efficiently extract the shape representation, which is usually formed by arranging the coordinates of fiducial facial points as a vector, is the key challenge and has significant influence on the overall performance of the algorithm.

This challenge of localizing fiducial facial points mainly comes from the complexity of appearance variations possibly exhibited in the patch centered at each facial point, caused by the change of lighting, pose, occlusion, expression and so on. The combination of these factors may lead to a very complicated nonlinear manifold of appearance which is hard to model. Actually, only a few salient facial components (e.g., eyes and nose) in our face can be stably localized due to their unique patterns, while others like feature points along one's face contour are difficult to locate using traditional image analysis methods simply because of too many degrees of freedom at these locations. To deal with these problems, numerous approaches have been proposed in recent decades. Some of them will be discussed in the next section.

In this paper we present a novel local hough voting method to improve the efficiency and accuracy of fiducial points localization, which can then be conveniently integrated with the standard global prior model under the Bayesian framework for final face alignment. The main idea of our method is to first localize very few stable facial components (e.g., eyes or nose) from the given face image, then use the locations of these to help reduce the ambiguity encountered when locating other less stable facial feature points. Improving accuracy of fiducial points localization progressively by incorporating the information of neighboring anchor points is shown to be effective in face alignment [3], [15], [24], [26]. For example, Valstar et al [24] build a hierarchical MRF using the locations of stable facial components to constrain the search space by exploiting the constellations that facial points can form. Liang et al [15] propose to adjust global shape based on facial components only. Cao et al [3] give a two-level cascaded regression for face alignment in which the location of stable facial components are used to guide the shape regression.

However, most of the above models are either complex in inference or needing many anchor points so as to provide reliable constraints. There exists the need to develop novel and more efficient way to incorporate the prior knowledge into the task of face alignment, and this is exactly the focus of this paper. In particular, our method is based on the Constrained Local Models (CLM, [6]) but are augmented with local Hough voting-based method to improve the results. Compared to others, our method is both simple and easy to implement. For example, only two stable facial components (e.g., the left and right eyes) are needed as anchor points in our method, compared to 7 in [24] and 11 in [15], which effectively helps to reduce the risk of possible mislocations when locating too many facial components and save the computational costs. In addition, this method is general and can be conveniently integrated in any CLM framework. Last but not least, we use the local spatial constraints imposed by the anchor facial components as prior knowledge to cast votes on the search space of each facial points instead of searching them aimlessly. Actually, in a deformable model method depending on an iterative optimizing procedure like Active Shape Model (ASM,[4]), properly initializing each local searching region is a very importance issue and inaccurate initialization could lead to a possibly very bad solution.

This local Hough voting strategy has close connection with the Implicit Shape Model (ISM,[13]) method, where each detected local part casts vote on the location of the object of interest. The ISM method has gained great success in the field of object detection. However, to our knowledge, it has not been used in face alignment. Furthermore, instead of directly searching the location most voted in the voting space, we combine the local voting map and the response map of local detectors using a multi-output ridge regression method,

X. Jin, X. Tan and L. Zhou are all with the Department of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China. {x.jin, x.tan}@nuaa.edu.cn zhouliang0806@sina.com

which effectively increases the robustness of the algorithm against various appearance variations. The non-parametric way to construct voting map in our method is also related to [1], where a large collection of diverse, labeled exemplars are used to provide the ”votes” and exemplars for best part localization in a holistic way. Finally, we also systematically investigate the impact of anchor points location accuracy, which is largely ignored in previous works. The above contributions of this paper are described in more detail in Section III and their effectiveness is verified in Section IV. Section V concludes our work.

II. Background

In this section we briefly introduce the basic face alignment framework within which some related work is discussed. A typical face alignment model involves three key components [4], [17], i.e., shape model, transformation model and image model. The shape model represents the geometric shape of the object of interest, the image model gives the physical constraints imposed on how the shape may vary over space, while the transform model bridges the ideal shape vector to the actual shape observed from the given image with a mathematical function. The fitting process are usually divided into two iterative stages [2], [4], [7]. In the first bottom up stage, a shape vector L is first extracted from the given face image I by detecting facial features using the image model, which are then undergone some geometric normalization with the transform model. The calibrated shape vector S is finally checked by the shape model. Besides evaluating the goodness of deformation for an input shape S, the shape model may recommend a new shape u, which is more consistent with the allowable deformation, to the system. The new shape u is eventually mapped to the image space along the same path but in reverse direction, and this process repeated until some converging condition is met. The main advantage of this methodology is that the ambiguity involved in localizing each facial features are largely reduced by the overall shape constraints imposed by the shape model.

According to this, existing face alignment methods are mainly different in their underlying assumptions involved in the construction of image model and shape model, while the transform model is usually taken to be linear for simplicity. Among them, the shape model plays the role of integrating prior shape constraints into the system. One of the most popular assumption about the shape model is the single mode Guassian distribution [4], which is only a very rough linear approximation to the global shape variations in faces. More realistic shape model is possible and adopted by the researchers, e.g., Mixture of Gaussian (GMM) [5], [15], Kernel principal component analysis (KPCA) [19]), Guassian Process Latent Variable Model (GPLVM) [11]), Markov Random Field (MRF) [9], [24]. In general, more complicated models are more closed to the real world but are usually coupled with higher cost for optimization at the same time. In addition, it becomes harder to ensure the goodness of generated shapes over a complex manifold. To overcome this, recently some authors advocate the use of non-parametric model [1], [3], [13] but a large number of diverse and labeled training samples are needed.

The image model is another key component of the face alignment system, which extracts geometric description of the shape from the given image. However, as mentioned in the previous section, this task is extremely challenging. To make things easier, instead of using a global likelihood model, many works in literatures adopt the strategy of divide and conquer, that is, constructing a local model for each fiducial point. One typical one is the Constrained local model (CLM) firstly proposed by D. Cristinacce and T. Cootes [6], in which the location of each fiducial point is searched using a sliding window method. Many pattern recognition techniques, either generative or discriminative, have been adopted to learn the local classifier by researchers: Adaboost [16], SVM [8], Mixture of linear SVMs [20], Convex quadratic fitting [25], Mixture of Gaussian [10], Mean-shift model [21], just to name a few.

It is worth mentioning that most of the previous CLM models assume that the local models are conditional independent given the face image. This improves the computational trackability but at the same time pushes too much burden of consistency checking for shape vector L extracted from the image to the back end shape model. However, one may conjecture that the stableness of the algorithm could be significantly improved if more reliable shape vector L could be obtained directly from the image instead of relying on the recommendation of the prior model. Following this, some researchers relax the independence assumption in CLM and introduce the local spatial constraints into the local model [14], [23], [24]. Our work in this paper belongs to this category of methods as well.

III. Local Hough Voting for Face Alignment

In this section, we describe how build our local voting space using a training image set and use it for the task of face alignment.

A. Overview

Given an image I, the task of face alignment is to locate M facial feature points $L = \{l_i = (x_i, y_i)\}_{i=1}^{M}$ on the 2D image, which are organized in a fixed order to a shape vector L. Denote the likelihood of generating this image by the facial parts at some locations $p(I|L, \theta)$, where $\theta = \{\theta_l, \theta_t, \theta_s\}$ is the parameters for the image model, transform model and shape model respectively. Assume that we are working on the region output by a face detector, and therefore we need not model the background. To infer the locations of the facial parts from this model, we look for the maximum *a posterior* $p(L|I, \theta)$, i.e., the probability that a face configuration is L given the model θ and an image I. According to Bayes rule, the posterior can be written as

$$p(L|I, \theta) \propto p(I|L, \theta_l)p(L|\theta_t, \theta_s), \tag{1}$$

where $p(I|L, \theta_l)$ is the generative image model of appearance while $p(L|\theta_t, \theta_s)$ is the shape model expresses the spatial constraints.

In this paper, we use the standard ASM formulation for shape prior $p(L|\theta_t, \theta_s)$, in which the shape vector L is assumed to be linearly transformed to a canonical position $S = \{s_i\}_{i=1}^{M}$. Denote $\theta_t = \{s, R, u_{xy}\}$, each fiducial point is then described by the following transform model,

$$l_i = sRs_i + u_{xy} \tag{2}$$

where s is the scale parameter, R is the rotation matrix, and u_{xy} is the centroid of the fiducial points. After this, S is the centralized and scale normalized fiducial points. Further assuming that the parameters of transform model is statistically independent with those of shape model, the final shape prior is given by,

$$p(L) = p(S|\theta_s)P(\theta_t) \tag{3}$$

As in ASM and many other points distribution models, the prior for canonical shape vector S is simply modeled as Gaussian and approximated using PCA in this paper, while the transformation model is separately learnt under the least square framework. We will describe our image model below.

B. Local Hough Voting

Let the image patch centered at each fiducial point be $T(l_i)$ and assume that the parts are statistically independent, the image model can be written to be,

$$p(I|L, \theta_l) = \prod_{i=1}^{M} p(T(l_i)|l_i, \theta_{l_i}), \tag{4}$$

By introducing an indicator variable y_i, the likelihood $p(T(l_i)|l_i, \theta_{l_i})$ of each part at location l_i can be evaluated using a discriminative model $p(y_i = 1|T(l_i), \theta_{l_i}, l_i)$, i.e., the probability that the appearance of patch $T(l_i)$ extracted at location l_i is just right, in other words, it is correctly aligned $(y_i = 1)$. A popular method to learn this model is the logistic regression,

$$p(y_i = 1|T(l_i), \theta_{l_i}, l_i) = \frac{1}{1 + e^{a\widehat{f}(T(l_i))+b}}, \tag{5}$$

where a, b is two parameters whose values can be obtained by cross validation, while the response of $\widehat{f}(T(l_i))$ is from a pre-trained SVM classifier, i.e., $\widehat{f}(T(l_i)) = \sum_k \alpha_k y_k T_k(i)' T(l_i) = T(l_i)' \sum_k \alpha_k y_k T_k(i)$, where $T_k(i)$ is the k-th support vector of the i-th fiducial point.

We note that although this method works well in some situations, the local patch representation restricts its generalization capability. A small patch brings about too much ambiguity while a larger one may suffer from too much appearance variations and can not capture the inherent characteristics of that fiducial point. Multi-scale descriptors help to alleviate this, while in this paper we choose to use local spatial constraints to reduce such ambiguity.

Specifically, given a test face image, we first localize K stable facial components (also called anchor points in this paper, e.g., the left and right eyes). Denote their locations as $\{p_j \in \Omega_j\}_{j=1}^{K}$, where Ω_j is spatial range each facial component may lie in, then the contribution of each anchor point j to the likelihood is,

$$p(y_i = 1|T(l_i), \theta_{l_i}, l_i) = \sum_{p_j \in \Omega_j} p(y_i = 1, p_j|T(l_i), \theta_{l_i}, l_i) \tag{6}$$

$$= \sum_{p_j \in \Omega_j} p(y_i = 1|T(l_i), \theta_{l_i}, l_i, p_j) p(p_j|T(l_i), l_i) \tag{7}$$

The first term is the probabilistic Hough vote by the j-th stable facial component at location p_j for the goodness of the i-th fiducial point given its patch representation. The second term specifies the spatial relationship between this facial component with the current fiducial point i. Note that this second term actually make it possible for us to improve the degree of tolerance of mislocations of anchor facial components, although we didn't do it in our current implementation. Instead, assuming that p_j is statistically independent with the appearance of patch $T(l_i)$, we have,

$$p(p_j|T(l_i), l_i) = p(p_j|l_i) \propto p(l_i|p_j)p(p_j) \tag{8}$$

where $p(p_j)$ gives the confidence of the j-th anchor point positioned at p_j. Assuming a very sharp distribution of $p(p_j)$ (i.e., unconfident locations of anchor points would not make much help in locating fiducial point i), and from E.q.(6),(8), we evaluate the score of the i-th fiducial point given the K pre-specified anchor points $\{p_j^*\}_{i=1}^{K}$ as follows,

$$score(l_i|p_1^*, p_2^*, ..., p_K^*) \propto \sum_{j=1}^{K} p(y_i = 1|T(l_i), \theta_{l_i}, l_i) p(l_i|p_j^*) \tag{9}$$

As described later, the first term produces the response map of local detectors, while the second term gives the voting map from each facial component (e.g., the left eye) to a candidate location of l_i, see Fig.1.

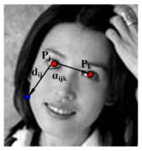


Fig. 1. The geometric constraints between three points (P_i and P_j are the anchor points).

C. *Non-parametric Smoothing for the Voting Map*

To cast the vote for the position l_i from the anchor facial component j, we need to model the distribution of $p(l_i|p_j^*)$. The vote can be understood as a prediction to the position of fiducial i after observing the position of anchor point j. However, since each face image is different to each other in different ways in scale, pose and lighting and so on, it will be difficult to directly model the position of each facial point with respect to the position of another point. Instead, we seek a more stable spatial constraints consisting of three points,i.e., $p(l_i|p_j^*) \approx p(l_i|p_j^*, p_k^*)$, where p_j^* and p_k^* are two pre-located anchor points (the left eye and the right eye here). This will greatly improve the robustness of the prediction against scale and in-plane rotation changes. Similar method is adopted in [22] and [24].

Specifically, for a triangle formed by the positions of two eyes and another point i, we model (see Fig.1): 1) the relative length d_{ij} between the interested facial point i and one of the anchor points j, and 2) the inner angle α_{ijk} between the edge $e_{i,j}$ and $e_{j,k}$. The relative length will be scaled by a factor of mean distance between two eyes over the training set. Note that these two quantities will keep invariant although the positions of each points could change with respect to various scales and in-plane rotation.

Given a test face image, we first locate the two eyes [1]. The next question is how to predict the location of another fiducial point i using this information. One apparent answer would be to find it directly from the training set. However, unless the exemplars are rich enough, we can seldom find a exact match of the positions of two eyes in the training set. One way to address this is to use a non-parametric density method to smooth the voting map. Denote $v_i = (d_{i,j}, \alpha_{ijk})$, we use a Gaussian kernel for this purpose,

$$p(l_i|p_j^*, p_k^*) = p(v_i) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{1/2}}exp\{-\frac{\|v_i - v_n\|^2}{2h^2}\} \tag{10}$$

where h is standard deviation of the Gaussian components (set to be 4 pixels in our experiments by cross validation) and v_n is the ground truth obtained from the training set.

D. *Fusing the Response and Voting Maps*

One problem remained is that even we search the location of fiducial point i at the most voted position in the voting map, we may still face the danger of missing the truth due to the complexity of appearance. To improve the robustness, it is useful to fuse the information from both the local spatial relationship and the response from local detector. As illustrated in Fig.2, fusing these helps to reduce the ambiguity in searching for the best response. For example, as shown in the second response map in the last row, a facial point located at the face contour has a wide range of response, but such kind of ambiguity is dealt well with the method by combining the information from the voting map and the response map (see the last map in the third row).

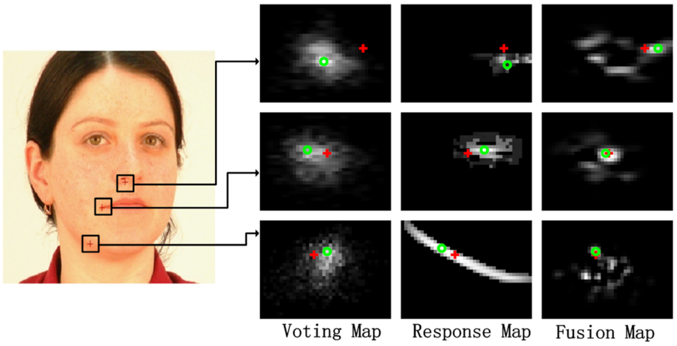


Fig. 2. Illustration of the voting map, response map and fused response map, where the red cross is the ground truth and the green circle is the position with maximum response of each map. This figure is best viewed in the electric form.

The central idea for this is to learn two linear transforms (rotations) for the voting map and response map respectively, such that after rotation both maps align well with the ground truth map available from the training set. Mathematically, denote the three maps of sample n as corresponding matrices V_n, E_n, and G_n, respectively. Note that they are normalized to the same size before performing the following steps. Then what we want to do is to learn two "rotation" matrices W_1 and W_2 for the voting map V_n and response map E_n respectively, such that the following criterion is satisfied,

$$\min_{W_1,W_2}\sum_{n=1}^{N}\|G_n-(W_1V_n+W_2E_n)\|_F^2+\lambda_1\|W_1\|_F^2+\lambda_2\|W_2\|_F^2 \tag{11}$$

where $\|\|_F$ is the Frobenius norm. In our implementation, we used a vector representation and concatenated the two maps into a single combined vector $u_n = [vec(V_n)^T, vec(E_n)^T]$. Further denote $W = [W_1, W_2]$ and $g_n = vec(G_n)$, then the above equation could be simplified as a standard multi-output ridge regression objective function,

$$\min_{W}\sum_{n=1}^{N}\|g_n - Wu_n\|_{l_2}^2 + \lambda\|W\|_F^2 \tag{12}$$

[1] We locate the positions of eyes using the method introduced in [22], and the sensitivity of our algorithm to the accuracy of eye localization is investigated in Section IV-B.2.

with the closed form solution,

$$W^{*} = (\sum_{n=1}^{N} g_n u_n^T)(\sum_{n=1}^{N} u_n u_n^T + \lambda I)^{-1} \qquad (13)$$

The regularization parameter λ is set to be a very small number (10^{-4} in our implementation).

IV. Experiments

To validate the effectiveness of the proposed method, we conducted a series of experiments on three publicly available databases with manually annotated landmarks, i.e., the PUT face database [12], the XM2VTS face database [18] and the BioID face database. The PUT database contains 9971 images from 100 subjects, and each image is annotated with a 194-point markup as ground truth landmarks. We used 1850 near-front faces of 100 subjects in our experiments, among which 1400 images from 70 subjects were used for model training while 450 images from the remaining 30 subjects for testing. For each image we choose 61 landmarks and add two landmarks of centers of eyes. Therefore, we have 63 points in all on each image (see Fig.3 (left)). The XM2VTS database consists of 2360 frontal face images of 295 subjects, with 68 ground truth annotations (but are different from the 63-point markup for PUT), see Fig.3 (middle). For this, 1760 images from 220 subjects were used for training and 600 images of the remaining 75 subjects for testing. The same settings are used for the ASM and CLM methods. On the other hand, in order to compare with some more advanced methods, we use the BioID database for experiment. The BioID database consists of 1521 gray level images with a resolution of 384x286 pixel. Each one annotated with 20 points shows the frontal view of a face of one out of 23 different test persons, see Fig.3 (right). For the BioID database, we use 1000 images of 18 different test persons for model training while 300 images from the remaining for testing.

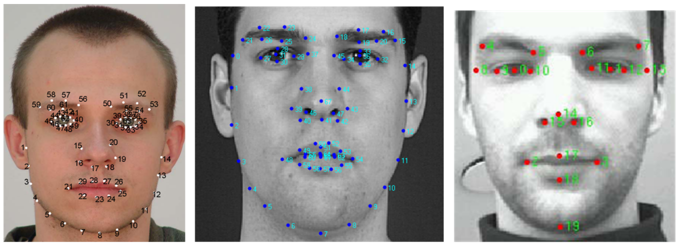


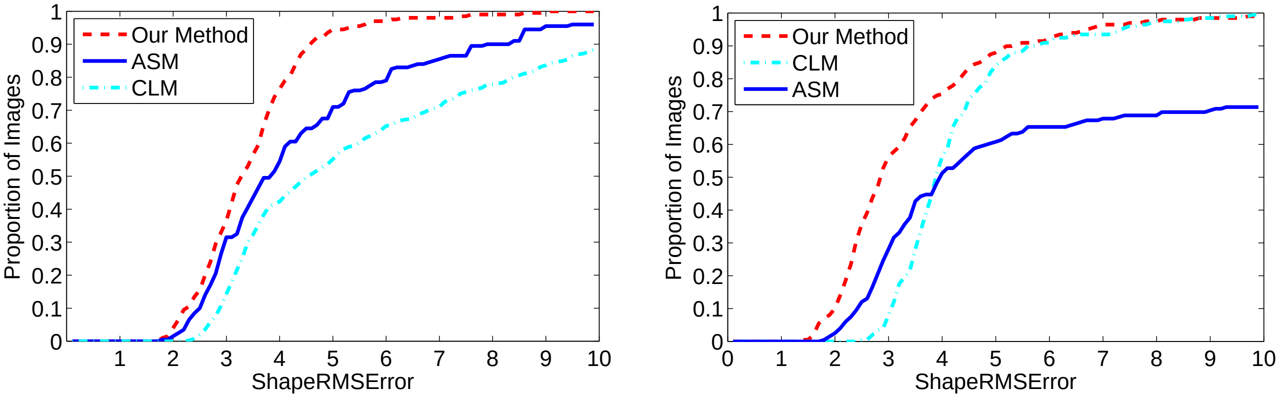
Fig. 3. Numbered landmarks for sample images from PUT (left), XM2VTS (middle) and BioID (right) .

A. Experiments on PUT and XM2VTS

Over these two databases we compared our method with ASM [4] and CLM [6]. In particular, we implemented a coarse to fine multi-scale matching strategy for ASM, in which at each lower fine level we refines the searching results from the higher coarse level (5 scales in all). For CLM, we extracted a patch of 11×11 centered at each landmark as positive sample while randomly sampling a few patches with the same size but located at least 8 pixels away from the center as negative samples. We simply used gray values as our feature descriptors and trained the detector using the SVM method with RBF kernel. In testing, the size of searching window for each facial point is 13×13 in pixels for both ASM and CLM, while for our method it is no need to set this since it is automatically defined by the voting map.

Fig.4 gives the fitting curves of our method compared with ASM and CLM on the PUT and XM2VTS databases, where the fitting curves show the proportion of images at various levels of overall localization errors, measured as the root-mean-squared (RMS) error between the ground truth landmarks and the resulting fit. The results show that our method consistently performs better than the compared methods on both databases, although the improvement on the XM2VTS database is marginal. In particular, on the PUT database, 94.5% of images have an RMS less than 5 pixels using our method, compared to 55.2% with CLM and 71.0% with ASM, while on the more realistic XM2VTS database, our method yields 87.9% of images at the same level of error, compared to 84.0% for CLM and 60.8% for ASM. Fig.6 shows some example alignment results using our method on the two databases. Note that none of these subjects appears in the training set.

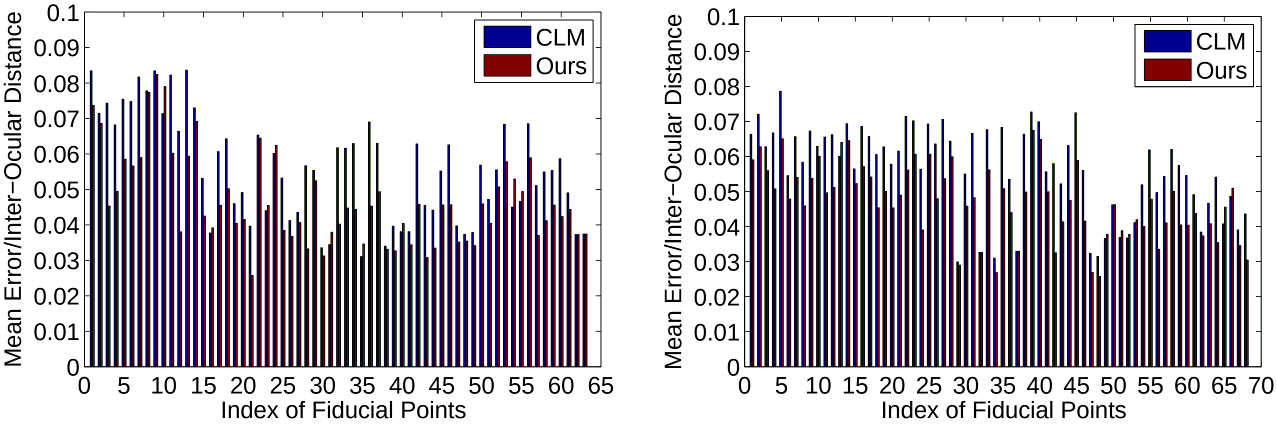
Fig.5 gives the detailed results of each local detector on the two databases, measured by the distance from each localized point to the ground truth (normalized with the inter-ocular distance). Note that for the PUT images, point number 62 and 63 denote the left eye and the right eye respectively and for XM2VTS images, the corresponding point numbers are 31 and 36 respectively (c.f., Fig.3). We can see from the figures that on both databases most of our detectors outperform their counterparts of CLM. In particular, 88.2% of our detectors have higher accuracy than those of CLM on the more challenging XM2VT database, while on the PUT database, about 85.7% detectors work better. By checking the distribution of those detectors with improved performance (c.f., Fig.3), we see that most of them are near the eyes, nose and the outer outline of the face. However, the improvements near mouth are less obvious, which is consistent with previous results [1], possibly due to the fact that the appearance of the mouth region can be easily influenced by the expression changes.



(a)

(b)

Fig. 4. Fitting curves for the ASM, CLM and the proposed method on the (a) PUT and (b) XM2VTS databases.



(a)

(b)

Fig. 5. Mean error of our fiducial detector on the PUT (left) and the XM2VTS (right) databases compared to the mean variation of CLM. The fiducial labels are shown in Fig.3. This figure is best viewed in the electric form.

B. *Experiments on BioID*

1) Comparison to other methods on BioID: Comparing the performance of face alignment algorithms is difficult in general, partly due to the lack of commonly adopted benchmark dataset. Fortunately, recently quite a few works have published their results on the BioID dataset, which allows us to perform a fair comparison with these methods. In particular, we compared our method with the results reported by [6], [24], [27], [28], see Fig.7. The figure shows that the overall performance using our local Hough voting method is slightly better than the method proposed by M. Valstar et al. [24] using hierarchical MRF model, both of which perform much better than extended ASM [27] and CLM [6]. Note that although Fig.7 shows that our method looks slightly inferior to the method of [1], the major goal of our method is to improve the accuracy of facial feature localization, which means that our method could be used to improve the performance of [1] in principle, as in the case of CLM.

2) The impact of anchor points accuracy: One possible criticism of the methods like ours lies in their dependance on the localization of anchor points, and hence investigating the impact of eye anchor points location accuracy is of interest. For this we disturb the ground truth locations of eyes randomly in each images by 1, 2, or 3 pixels, respectively, then evaluate the corresponding performance by calculating the fraction of landmarks with offset less than 5, 7.5, 10 pixels, respectively. We compared these with the results using the ground truth eye locations, and those using the output of the automatic eye detector [22] adopted in this paper.

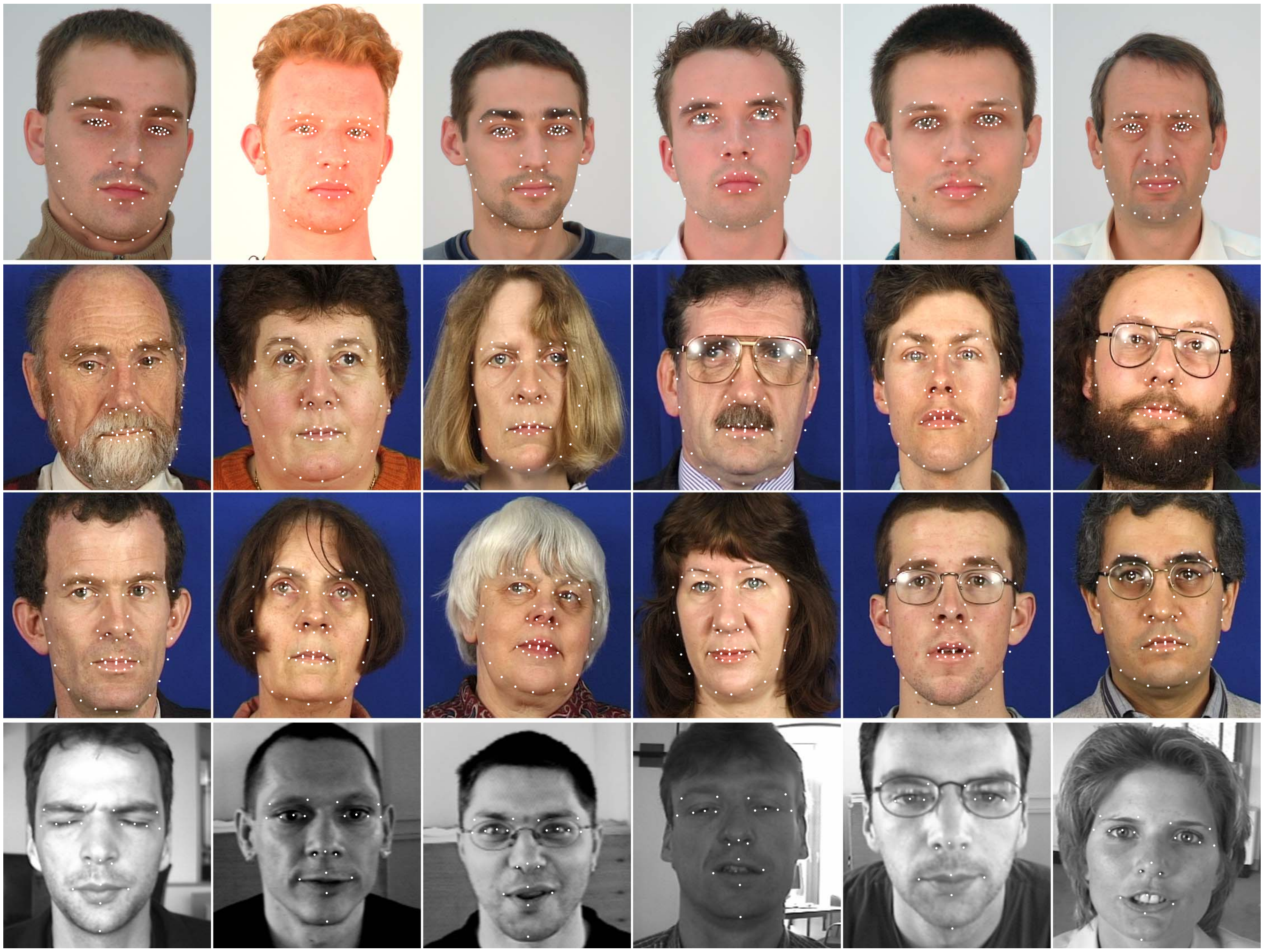


Fig. 6. Illustration of the aligned images from the PUT (top row), XM2VTS (middle two rows) and BioID (bottom row) using the proposed method.

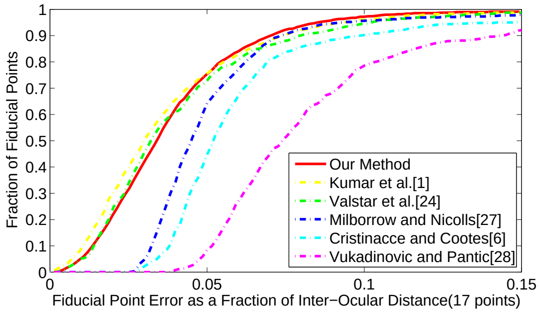


Fig. 7. Cumulative error distribution curves comparing our method to several others on the BioID database. For a fair comparison with previous results, only 17 landmarks are used.

Fig.8 gives the overall comparative results, which indicates that the proposed face alignment algorithm is a bit sensitive to the random noise of the locations of anchor points, especially when the eye locations are extremely unreliable (3 pixels away from the ground truth). This is as expected, since we actually performed a worst-case testing by disturbing the ground truth consistently and separately on each image, which essentially disorders the statistical regularity encoded in the training data at least to some degree. Despite this, the figure shows that current eye detector works fairly well in this case. Indeed, Fig.8 reveals that the results obtained using the output of an automatic eye detector are only slightly worse than those using ground truth. We further examined the localization accuracy of this eye detector and found that in most cases it yields localization error in less than 1-2 pixels although it is difficult to precisely locate eyes in some images (c.f., Fig.6). Even so, it still leads to better alignment performance than randomly disturbing the ground truth eye position with only 1 pixel (c.f., Fig.8). This clearly indicates that the feasibility of the proposed method in practice.

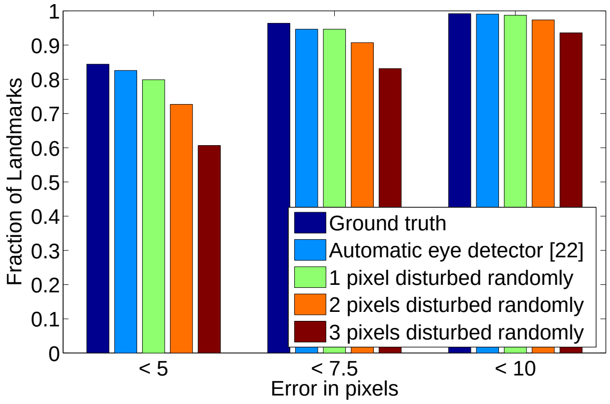


Fig. 8. Impact of eye localization accuracy on the performance.

V. Conclusion

In this paper we propose a novel local Hough voting based method for face alignment. The key idea of this method is to use the spatial constraints imposed by the stable facial components to guide the search of other facial points. We show in this paper that it is possible to implement this by first constructing a voting map with non-parametric kernel smoothing, then fusing it with the traditional response map using the multi-output ridge regression method. Compared to others, the proposed method is efficient, simple to implement and can be conveniently integrated with any prior shape model and more powerful individual local models. We demonstrate its effectiveness on several publicly available databases with promising results.

Acknowledgment

The work was financed by the (key) National Science Foundation of China (61073112, 61035003), and the National Science Foundation of Jiangsu Province (BK2012793).

References

[1] Peter N. Belhumeur, David W.Jacobs, David J. Kriegman, and Neeraj Kumar.Localizing parts of faces using a consensus of exemplars.*In CVPR*, June 2011.

[2] M.C. Burl,T.k. Leung,and P. Perona. Face localization via shape statistics.*In AFGR*,1995.

[3] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression.*In CVPR*, 2012.

[4] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, et al. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995

[5] Timothy F. Cootes and Christopher J. Taylor. A mixture model for representing shape variation. *Image Vision Comput.*, 17(8):567–573, 1999.

[6] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. *In BMVC*, volume 3, pages 929–938, 2006.

[7] Liya Ding and Aleix M. Martinez. Precise detailed detection of faces and facial fea- tures. *In CVPR*, volume 0, pages 1–7, 2008.

[8] L. Ellis, N.D.H. Dowson, J.G. Matas, and R. Bowden. Linear predictors for fast simul- taneous modeling and tracking. *In NRTL07*, pages 1–8, 2007.

[9] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recog- nition. *Int. J. Comput. Vision*, 61(1):55–79, 2005.

[10] L. Gu and T. Kanade. A generative shape regularization model for robust face align- ment. *In ECCV*, pages I: 413-C426, 2008.

[11] Y.C. Huang, Q.S. Liu, and D.N. Metaxas. A component based deformable model for generalized face alignment. *In ICCV*, 2007.

[12] A. Kasinski, A. Florek, and A. Schmidt. The put face database. *Image Processing & Communications*, 13(3–4):59–64, 2008.

[13] Bastian Leibe, Ale Leonardis, and Bernt Schiele. Robust object detection with interleaved categorization and segmentation. *Int. J. Comput. Vision*, 77:259–289, 2008.

[14] L. Liang, F. Wen, Y.Q. Xu, X. Tang, and H.Y. Shum. Accurate face alignment using shape constrained markov network. *In CVPR*, pages I: 1313–1319, 2006.

[15] Lin Liang, Rong Xiao, Fang Wen, and Jian Sun. Face alignment via component-based discriminative search. *In ECCV*,pages 72–85, 2008.

[16] X.M. Liu. Discriminative face alignment. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 31(11):1941–1954, November 2009.

[17] T. McInerney and D. Terzopoulos. Deformable models in medical image analysis: A survey. *Medical Image Analysis*, 1(2):91–108, 1996.

[18] K. Messer, J. Matas, J. Kittler, and K. Jonsson. XM2VTSDB: The extended XM2VTS database. *In AVBPA*, pages 72–77, 1999.

[19] Romdhani S., S. Gong, and A. Psarrou. Multi-view nonlinear active shape model using kernel PCA. *In BMVC*, 1999.

[20] Jason M. Saragih, Simon Lucey, and Jeffrey Cohn. Deformable model fitting with a mixture of local experts. *In ICCV*, 2009.

[21] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Deformable model fitting by regularized landmark mean-shift. *Int. J. Comput. Vision*, 91(2):200–215, January 2011. ISSN 0920–5691.

[22] X. Tan, F. Song, Z.H. Zhou, and S. Chen. Enhanced pictorial structures for precise eye localization under incontrolled conditions. *In CVPR*, pages 1621–1628. IEEE, 2009.

[23] P.A. Tresadern, H. Bhaskar, S. Adeshina, C.J. Taylor, and T.F. Cootes. Combining local and global shape models for deformable object matching. *In BMVC09*, 2009.

[24] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. *In CVPR*, pages 2729–2736. IEEE, 2010.

[25] Y. Wang, S. Lucey, and J.F. Cohn. Enforcing convexity for improved alignment with constrained local models. *In CVPR*, pages 1–8, 2008.

[26] Tiddeman, B. Facial feature detection with 3D convex local models. *In FG*, pages 400–405. IEEE, 2011.

[27] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *In ECCV*, pages 504–513, 2008. 546, 550

[28] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. *In ICSMC*, 2:1692–1698, 2005.