FACE ALIGNMENT BASED ON THE MULTI-SCALE LOCAL FEATURES

Cong Geng, Xudong Jiang

Nanyang Technological University
Electrical and Electronic Engineering
Nanyang Link, Singapore 639798

ABSTRACT

Many face recognition algorithms depend on careful positioning of face images into the same canonical pose. Currently, this positioning is usually done by detecting the locations of eyes. And the face images are transformed to the same positions according to the eye coordinates detected. In this paper, we describe a method based on multi-scale local features to achieve face alignment automatically not just dependent on the localizations of two eyes. Given an unaligned face image resulting from a face detector and a set of aligned face images in the data set, we build an automatic transformation mechanism, under which the unaligned face image can be precisely aligned for the following recognition process. Our alignment method improves performance on face recognition tasks, over images aligned by many other algorithms.

Index Terms— face alignment, multi-scale local features, eye detection, face recognition

1. INTRODUCTION

Since the Principal Component Analysis (PCA) [1] and the Linear Discriminant Analysis (LDA) [2] were introduced into face recognition, various holistic approaches have been extensively studied [3]. However, the holistic approaches require a preprocessing procedure to normalize the face image variations in pose and scale, which is not an easy task because it depends on the accurate detection of at least two landmarks from the face image. Some algorithms for eye localization have been proposed based on the eyeball [4, 5, 6, 7, 8]. However, in many real applications the appearances of eyeball are not distinct or missing due to expressions, occlusions, illuminations or image noise. Hence, some algorithms localize multiple facial features like corners of eyes, nostrils, the tip of nose, corners of mouth, etc. Face alignment is performed based on these semantic features [9]. The same problem encountered in the detection of eyes remains. Moreover, in the training process, these semantic features are hand-annotated, which is very labor-consuming. In [10], an unsupervised approach is proposed for face alignment, which is not based on the localizations of semantic facial features. As the performance of the face alignment algorithm influences the final recognition performance, many research papers on the holistic approaches report the recognition performance on the pre-normalized faces. The recognition performance will deteriorate considerably if the manual process is replaced by an automatic landmark detection algorithm.

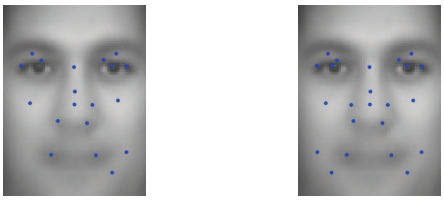
In contrast to holistic methods, some local feature based approaches for face recognition are more robust to variations in pose and scale. Furthermore, unlike the holistic approaches, the face normalization is an integrated part of the local approaches [11, 12, 13, 14]. To solve the alignment problem in holistic approaches, we propose a face alignment strategy based on multi-scale local features instead of just two specific eye points. In [15], a method for partial face alignment in near infrared (NIR) video sequences is proposed based on SIFT [11]. Different from this approach [15], the anchor points in our template face image are detected and learned automatically. In the alignment stage, we do not use shape constraint [15] which is limited to align frontal faces with slight pose variations. Instead, we use Hough transform to cluster keypoints with similar poses and then apply affine transform to each cluster to remove spurious correspondences. In this way, we can align faces with large pose variations. The performance of our face alignment strategy is validated by face recognition tasks using holistic approaches LDA [2], UFS [16] and ERE [17]. Experimental results on Georgia Tech (GT) [18] and ORL [19] databases show that our alignment approach outperforms those based on localization of eyes [4, 5, 6, 7, 8], the localization of facial parts [9] and the congealing approach [10].

2. FACE ALIGNMENT

The purpose of our alignment is to rectify face images into the same canonical pose for subsequent holistic recognition tasks, rather than localizing facial feature points such as eye-brows, eyes, nose, mouth and contour of chin as many papers did. As mentioned in Section 1, face alignment algorithms based on localizations of facial parts are not reliable as the appearances of semantic facial features vary with expressions, illuminations, occlusions or image noise. Hence, we propose an approach for face alignment not just relying on the semantic facial parts.

2.1. Generate the Common Face Template

Given a set of face images $\mathcal{O}$ in the training database, we align them in pose and scale with manually detected two eye coordinates. Our goal in this step is to learn a common face template based on these aligned face images $\mathcal{I}$. As mentioned above, the similarities are high, among the facial componential appearances of different subjects. The mean face $\mathfrak{m}$ of $\mathcal{I}$ captures the common information of various identities and removes noises. The SIFT keypoints detected in $\mathfrak{m}$ tell us the locations of the common and stable features in $\mathcal{I}$. Fig. 1(a) shows the mean face $\mathfrak{m}$ computed from $\mathcal{I}$ with SIFT keypoints. We further add extra keypoints to meet the symmetry property of face images, as shown in Fig. 1(b), where the keypoints are the anchor points in the common face template $\mathfrak{m}$.



(a) (b)

Fig. 1: Mean face $\mathfrak{m}$ with (a) SIFT keypoints; (b) Symmetric keypoints $\mathcal{K}$.

The anchor points in $\mathfrak{m}$ tell us the possible locations of common features in $\mathcal{I}$. However, their descriptors provide little information, as $\mathfrak{m}$ is the mean face image. The support area of SIFT descriptors in $\mathfrak{m}$ is smoothed by the mean. We need to compute the descriptors directly from the individual images in the original training set $\mathcal{O}$. As there are pose variations, the locations of the detected keypoints in $\mathcal{O}$ cannot be used in the alignment process. We should project their locations into the coordinates of the well-aligned image set $\mathcal{I}$. Let $\mathcal{P}_q = \{p_i^q\}$, $q = 1, \ldots Q$, where Q is the number of images in the training set $\mathcal{O}$, represent the keypoint set detected in the q^{th} image of $\mathcal{O}$, where p_i^q is the i^{th} keypoint in $\mathcal{P}_q$. Suppose that the two eye coordinates of the q^{th} face image in the set $\mathcal{O}$ are $[e_{1x}, e_{1y}; e_{2x}, e_{2y}]$, and the two eye coordinates in the corresponding well-aligned face image in the set $\mathcal{I}$ are $[a_{1x}, a_{1y}; a_{2x}, a_{2y}]$. Based on these two pairs of corresponding points, we can compute the similarity transformation parameters $[s, \theta, t_x, t_y]$ between the q^{th} image in $\mathcal{O}$ and the well-aligned image in $\mathcal{I}$ as below:

$$\begin{bmatrix} s\cos\theta \\ s\sin\theta \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} e_{1x} & -e_{1y} & 1 & 0 \\ e_{1y} & e_{1x} & 0 & 1 \\ e_{2x} & -e_{2y} & 1 & 0 \\ e_{2y} & e_{2x} & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} a_{1x} \\ a_{1y} \\ a_{2x} \\ a_{2y} \end{bmatrix} \quad (1)$$

Once we get the transformation parameters, we can project the location of the keypoint p_i^q, $[x_p, y_p]$, to the corresponding well-aligned coordinates $[x'_p, y'_p]$ by

$$\begin{bmatrix} x'_p \\ y'_p \end{bmatrix} = \begin{bmatrix} s\cos\theta & -s\sin\theta \\ s\sin\theta & s\cos\theta \end{bmatrix} \begin{bmatrix} x_p \\ y_p \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \quad (2)$$

We perform this projection on the locations of each keypoint in the set $\mathcal{P}$. Thus the keypoint descriptors of $\mathcal{P}$ capture various pose information, and their locations are well-aligned.

Let $\mathcal{K} = \{k_i\}$, $i = 1, \ldots t$, denote the anchor point set in $\mathfrak{m}$, where t is the number of anchor points in the common face template. Image set $\mathcal{O}$ is the original images with pose variations. The keypoint descriptors of $\mathcal{P}$ capture various pose information. To enhance the representative power of the template image $\mathfrak{m}$, we embed the descriptors of $\mathcal{P}$ into the anchor keypoint set $\mathcal{K}$. In a region R around the location of k_i, we search its neighbors in $\{\mathcal{P}_1, \ldots, \mathcal{P}_Q\}$. R is set to $1/6$ times the image size in the experiments. If there are multiple keypoints in one face image falling into R, we select the one which is nearest to the location of k_i. In this way, around each anchor point k_i, we can find a series of keypoints $\mathcal{N}_i$ from different face images. Comparing with the scheme of one anchor point with one descriptor from one face image, variance of keypoints coming from the same sematic region of different faces enrich feature representation and are less subject to pose variations. Note that we only use the location information of anchor point k_i to locate candidate keypoints nearby in $\mathcal{P}$. We do not use the location or the descriptor of the anchor point k_i during the alignment process.

Now around each anchor point k_i, there is a series of keypoints $\mathcal{N}_i$. To make the number of keypoints in $\mathcal{N}_i$ less dependent to the number of images Q in the training database $\mathcal{O}$, we adopt hierarchical clustering [20] to group the descriptors of each keypoint set $\mathcal{N}_i$ into h clusters. The cluster center c_j^i, where $j = 1, \ldots, h$, is selected as the descriptor who has the largest accumulated cosine similarities among all the other descriptors in the same cluster. If the number of keypoints in $\mathcal{N}_i$ is smaller than h, we keep all the keypoints in $\mathcal{N}_i$. Hence, in the common face template $\mathfrak{m}$, the final number of keypoints is smaller than or equal to $t \times h$. And we denote these keypoints in the template image $\mathfrak{m}$ as final anchor point set $\mathcal{T}$.

2.2. Establish the Feature Correspondences

Now in the template image $\mathfrak{m}$, there are at most $t \times h$ anchor points extracted from various face images. Suppose that image *I* is the output of some face detector, which should be aligned into the same canonical pose as the template image $\mathfrak{m}$ for the subsequent holistic recognition process. SIFT keypoint set $\mathcal{B}$ is extracted from the image *I*. The best candidate match of a probe keypoint in $\mathcal{B}$ is found by identifying its nearest neighbor in the anchor point set $\mathcal{T}$. The nearest neighbor is defined as the anchor point whose descriptor has the maximum similarity to that of the probe keypoint.

The nearest-neighbor search can only establish putative correspondences between keypoint sets $\mathcal{B}$ and $\mathcal{T}$. To elimi-

(a)

(b)

Fig. 2: Sample images (a) before alignment; (b) after alignment.

nate spurious keypoint pairs, we further check their geometric consistencies by the Hough transform and following the affine transformation described in [11]. After the geometric verification, we can obtain keypoint pairs $\mathcal{B}_{sub}$ and $\mathcal{T}_{sub}$, which are subsets of $\mathcal{B}$ and $\mathcal{T}$ respectively. Note that some putative keypoint pairs are rejected by this process due to their geometric inconsistency. The anchor point set $\mathcal{T}_{sub}$ contains the final anchor points for the keypoint set $\mathcal{B}_{sub}$ to align to.

2.3. Face Alignment by Similarity Transformation

The purpose of our face alignment is to rotate, resize and crop the output face images of face detectors automatically, which transforms them into canonical pose for the subsequent holistic recognition tasks. We do not want to change their structures. Hence, we adopt similarity transformation in the final alignment step. The similarity transformation gives the mapping of a model point $[x, y]$ to an image point $[u, v]$ in terms of an image scaling s, rotation θ, and translation $[t_x, t_y]$. We project the location of a probe keypoint $[x_p, y_p]$ in $\mathcal{B}_{sub}$ to its corresponding anchor keypoint $[x_p{}', y_p{}']$ in $\mathcal{T}_{sub}$ by the similarity transform as

$$\begin{bmatrix} x_p & -y_p & 1 & 0 \\ y_p & x_p & 0 & 1 \\ & \cdots & & \\ & \cdots & & \end{bmatrix} \begin{bmatrix} s\cos\theta \\ s\sin\theta \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} x_p{}' \\ y_p{}' \\ \cdot \\ \cdot \end{bmatrix} \quad (3)$$

Each matched keypoint pair contributes two rows to the first and last matrices in Eq. 3. At least 2 matches are needed to provide a solution. We can write this linear system as

$$\mathbf{Ax} = \mathbf{b} \quad (4)$$

The least-squares solution for the parameters **x** can be determined by solving the corresponding normal equations

$$\mathbf{x} = [\mathbf{A^T A}]^{-1}\mathbf{A^T b}, \quad (5)$$

which minimizes the sum of the squares of the distances from the projected model locations to the corresponding image locations.

Once we obtain the transformation parameters [s, θ, t_x, t_y], we can transform the probe image I according to the 2-D spatial similarity transformation. Fig. 2(a) shows some sample images before alignment. Fig. 2(b) shows the corresponding images aligned by our approach.

3. EXPERIMENTAL RESULTS

Two databases, GT [18] and ORL [19] are applied to test the face alignment performance through three face recognition algorithms LDA [2], UFS [16] and ERE [17]. Eight face alignment approaches are compared: Adaboost eye detector [4, 5], eye localization by pixel differences (PD) [6], eye localization by rank order filter (ROF) [7], eye localization by cascaded asymmetric principal and discriminative analysis (C-APCDA) [8], localization of semantic facial features(SF) [9], Congealing [10], our proposed approach (Prop.) and manual alignment (MA).

Table 1: Recognition rate on GT and ORL databases

GT			
	LDA	UFS	ERE
MA	92.00%	91.43%	92.86%
Adaboost [4, 5]	82.29%	81.71%	83.71%
PD [6]	64.29%	69.14%	71.43%
ROF [7]	60.29%	64.00%	67.43%
C-APCDA [8]	83.43%	83.71%	86.29%
Congealing [10]	84.29%	82.29%	85.71%
SF [9]	83.43%	84.57%	86.00%
Prop.	**92.57%**	**90.57%**	**92.00%**
ORL			
	LDA	UFS	ERE
MA	92.5%	83.5%	97.0%
Adaboost [4, 5]	88.5%	77.0%	91.0%
PD [6]	80.5%	69.5%	84.5%
ROF [7]	74.5%	65.5%	80.0%
C-APCDA [8]	83.0%	62.5%	88.0%
SF [9]	93.0%	81.5%	96.5%
Prop.	**93.0%**	**83.0%**	**95.5%**

The images of GT database have large variations in both pose and expression and some illumination changes. They are converted to gray scale and cropped into the size of 60×80. At the alignment stage, for C-APCDA [8] and our proposed approach, the first 8 images per subject serve as training images and are aligned manually by two eye coordinates. The remaining 7 images per subject serve as the output of the face detector, which should be aligned. At the recognition stage, for the three holistic approaches LDA [2], UFS [16] and ERE [17], the first 8 images of all subjects are used in the training and gallery sets, which are normalized manually. The remaining 7 images of all subjects serve as probe, which are aligned by different approaches. Images of ORL database are cropped into the size of 50×57. The first 5 images per subject serve

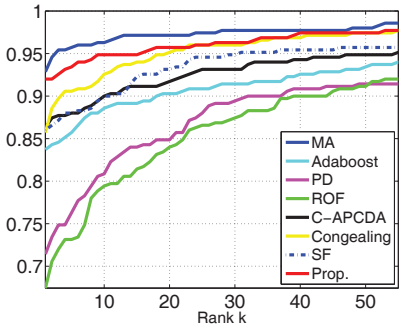


Fig. 3: Cumulative matching curves of eight different alignment approaches obtained by ERE on GT database (best viewed in color).

as training images. The remaining images serve as the output of the face detector. The best recognition performances of the holistic approaches LDA [2], UFS [16] and ERE [17] over all possible numbers of features are recorded.

Table 1 shows the rank one recognition rates on GT and ORL databases of three recognition algorithms based on different alignment approaches. On both databases, comparing different automatic face alignment approaches, the performance of our face alignment approach is significantly better than other algorithms. Fig. 3 gives the cumulative matching curves of GT database obtained from eight different alignment approaches based on the ERE approach. The cumulative recognition performance obtained by our alignment approach is significantly better than other automatic alignment approaches. And our alignment approach can achieve comparable results as those obtained by manual alignment.

4. CONCLUSION

Many face recognition algorithms depend on careful positioning of the face images into the same canonical pose, such as holistic approaches PCA, LDA and their variants. This is not an easy task because it depends on the accurate detection of at least two landmarks from the face image. Some algorithms for eye localization have been proposed based on the eyeball. And some algorithms are based on the localizations of semantic facial parts like corners of eyes, nostrils, corners of mouth, etc. However, in many real applications the appearances of these semantic features are not distinct or missing due to expressions, occlusions, illuminations or image noise, which makes the alignment results unreliable. To solve this problem, this paper presents a face alignment strategy based on multi-scale local features rather than semantic facial parts.

Given a set of aligned images in the data set, we firstly learn a face template in which the keypoints are from various aligned images. Then putative correspondences are established between the keypoint sets from the unaligned face image and the learnt face template by nearest neighbor search. Geometric verifications are performed to eliminate spurious matches with inconsistent poses. Then we build a transformation mechanism based on the final corresponding keypoint pairs, under which the unaligned face image can be precisely aligned with the template. The alignment process is not dependent on the localizations of facial parts, which leads to more reliable results.

5. REFERENCES

[1] M. Kirby and L. Sirovich, "Application of karhunen-loeve procedure for the characterization of human faces," *PAMI*, vol. 12, no. 1, pp. 103–108, January 1990.

[2] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *PAMI*, vol. 19, no. 7, pp. 711–720, July 1997.

[3] X.D. Jiang, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16–26, March 2011.

[4] A. Hadid, J.Y. Heikkila, O. Silver, and M. Pietikäinen, "Face and eye detection for person suthentication in mobile phones," *IEEE International Conference on Distributed Smart Cameras*, pp. 101–108, 2007.

[5] S.U. Jung, Y.S. Chung, J.H. Yoo, and K.Y. Moon, "Real-time face verification for mobile platforms," *Advances in visual computing*, pp. 823–832, 2008.

[6] J.F. Ren and X.D. Jiang, "Fast eye localization based on pixel differences," *IEEE Int. Conf. on Image Processing*, pp. 2733–2736, 2009.

[7] J.F. Ren and X.D. Jiang, "Eye detection based on rank order filter," *IEEE Int. Conf. on Information, Communication and Signal Processing*, pp. 1–4, Dec 2009.

[8] J.F. Ren, X.D. Jiang, and J. Yuan, "A fast and accurate cascade subspace face/eye detector on mobile devices," *Int. Conf. on Computer Vision, Workshop on Mobile Vision*, Nov. 2011.

[9] M. Everingham, J. Sivic, and A. Zisserman, ""Hello! My name is... Buffy" – automatic naming of characters in TV video," *BMVC*, pp. 889–908, Sep. 2006.

[10] G.B. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images," *ICCV*, pp. 1–8, Dec. 2007.

[11] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.

[12] C. Geng and X.D. Jiang, "Face recognition using sift features," *IEEE Int. Conf. on Image Processing*, pp. 3313–3316, Nov. 2009.

[13] C. Geng and X.D. Jiang, "Sift features for face recognition," *IEEE Int. Conf. on Computer Science and Information Technology*, pp. 598–602, Aug. 2009.

[14] C. Geng and X.D. Jiang, "Face recognition based on the multi-scale local image structures," *Pattern Recognition*, vol. 44, no. 10-11, pp. 2565–2575, Oct.-Nov. 2011.

[15] J.M. Yang, S.C. Liao, and S.Z. Li, "Automatic partial face alignment in nir video sequences," in *Proceedings of the Third International Conference on Advances in Biometrics*, 2009, pp. 249–258.

[16] X. Wang and X. Tang, "A unified framework for subspace face recognition," *PAMI*, vol. 26, no. 9, pp. 1222–1228, Sep. 2004.

[17] X.D. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 383–394, March 2008.

[18] A.V. Nefian, "Embedded bayesian networks for face recognition," *ICME*, pp. 133–136, Aug. 2002.

[19] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," *2nd IEEE Workshop on Applications of Computer Vision*, pp. 138–142, Dec. 1994.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer, 2009.