Automatic Partial Face Alignment in NIR Video Sequences

Jimei Yang[1,2], Shengcai Liao[2], Stan Z. Li[2]

[1] University of Science and Technology of China, Hefei 230027, China
[2] Center for Biometrics and Security Research, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{jmyang,szli}@cbsr.ia.ac.cn
http://www.cbsr.ia.ac.cn

Abstract. Face recognition with partial face images is an important problem in face biometrics. The necessity can arise in not so constrained environment such as in surveillance video or portal video (eg. as provided in Multiple Biometrics Grand Challenge (MBGC)). Face alignment with partial face images is a key step toward this challenging problem.
In this paper, we present a method for partial face alignment based on the well-known scale invariant feature transform (SIFT). We first train a reference model using holistic faces, in which the anchor points and their corresponding descriptor subspaces are learned from initial SIFT keypoints and the relationships between the anchor points are also derived. In the alignment stage, correspondences between the learned holistic face model and an input partial face image are established by matching keypoints of the partial face to the anchor points of the learned face model. Furthermore, shape constraint is used to eliminate outlier correspondences and temporal constraint is explored to find more inliers. Alignment is finally accomplished by solving a similarity transform. Experiments on the MBGC near infrared video sequences show the effectiveness of the proposed method, especially when PCA subspace, shape and temporal constraint are utilized.

Key words: Face Alignment, Partial Faces, SIFT, MBGC

1 Introduction

Face recognition is an important problem in both computer vision and biometrics. Most of researchers historically dealt with faces under constrained circumstances. However, with the development of state-of-the-art, researchers are shifting their interest to less constrained circumstances such as in surveillance video or portal video, where partial face recognition becomes a new challenge. Face alignment with partial face images is a prerequisite for solving this problem. Popular face alignment methods are mostly based on holistic face model, such as Active Shape Model (ASM)[1], Active Appearance Model (AAM)[2]. When the integrality of faces cannot be guaranteed, these holistic models will lose their

power. In this paper, we present a method for automatic partial face alignment based on scale invariant feature transform (SIFT)[3].

As a well-known local feature, SIFT has been used to perform face detection[4]. When performing in faces, SIFT keypoints have good repeatability in the same semantic region of different faces. By this property a set of uniform keypoints, called facial anchors, can be learned from a training database of holistic frontal faces. Descriptors attached to the same facial anchor together form its description. All of facial anchors and their respective descriptions compose a face model. When dealing with novel partial faces, we can establish point-wise correspondences between novel faces and the face model by matching keypoints to facial anchors. This sort of point-wise correspondences guarantee alignment robust to partial faces. Like the method in [3], we use the ratio of first best match and second best match to estimate the correctness of this correspondence. Some mismatches, however, still exist. Carneiro and Jepson[5] use shape context as semi-local feature integrated into SIFT to improve matching performance. Shape context[6] was originally proposed by Serge Belongie et al. to describe the object shape by shape point orientation histogram. In our setting, as far as sparsity is concerned, the orientation histogram of facial keypoints is not a stable feature so that we directly use shape constraint to prune outlier correspondences. A similarity transform can be solved from valid correspondences by using the method proposed in [7]. However, the number of valid correspondences established by one image is limited, which probably increase the risk of incorrect alignment. Temporal constraint in a video is further explored to enrich the pool of inlier correspondences. Two implications are derived from temporal constraint: pose continuity and identity consistency, both of which make it presumably easy to align faces to the same pose within a video. As a consequence, each of faces can contribute its inlier correspondences to one common similarity transform. It improves greatly alignment's robustness.

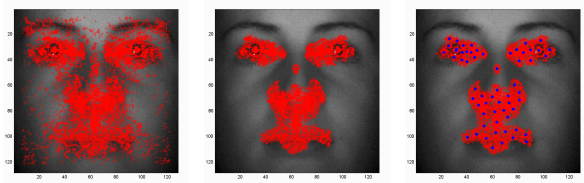
Our main contributions include: 1)SIFT based generative face model is learned and naturally overcomes the difficulties brought by face incompletion. 2)Shape constraint of face is used to prune correspondences, which guarantees that true matches are preserved and meanwhile most of false matches are eliminated. 3)Our method takes advantage of temporal constraint within a video to enrich correspondence inliers and thus obtain more robust results than frame-by-frame alignment.

The rest of this paper is organized as follows. In Section 2, a generative face model is built up. In section 3, partial face alignment with shape constraint and with temporal constraint are introduced in detail. In section 4, experiments are conducted on NIR video sequences released by Multiple Biometrics Grand Challenge (MBGC)[8] and show the performance of our method. This paper is finally concluded by discussion and future work in section 5.

2 Learning Face Model with SIFT

Scale Invariant Feature Transform (SIFT) developed by Lowe[3] combines a scale invariant keypoint detector and gradient histogram based descriptor. First, image scale-space is built and potential keypoints are identified by using difference of Gaussian (DoG) function. Final keypoints are selected based on measures of their stability. DoG as a close approximation of scale normalized Laplacian of Gaussian performed well in term of detection repeatability compared with other existing detectors[9]. The stability provides a prime basis for our face model. Second, the local coordinate system of keypoint is built in image scale-space. Based on the local coordinate system, The descriptor is represented by a 3D histogram of gradient locations and orientations. The contribution to the location and orientation bins is weighted by the gradient magnitude. The quantization of locations and orientations makes the descriptor robust to small geometric distortions and small errors in the keypoint detection. Mikolajczyk and Schimid[10] compared diverse existing descriptors and found SIFT best.

Using a set of holistic frontal faces as training database, we collect their SIFT keypoints and corresponding descriptors. All the keypoints are plotted in one typical face in Fig.1(a). Note that most of keypoints concentrate into small clusters around semantic facial regions such as eyes, nose and mouth. Considering location errors of DoG detector and feature displacements of different faces, we presume that each identifiable cluster of keypoints represent a facial semantic region and the mean of keypoints will be a good estimate of facial anchor. On the contrary, dispersive keypoints tend to be subject to some special features of certain faces and lack of generality and are thus removed before identifying facial anchors. We use following algorithm to remove dispersive keypoints: a keypoint, if the number of its neighbors within a small region R is less than certain number N_n, is considered dispersive and then removed. In our experiments, we set $R = 5$ and $N_n = 100$. After checking all keypoints, the remaining are shown in Fig.1(b). Finally, we identify facial anchors using Kmeans, which are represented by blue dots in Fig.1(c). Each facial anchor corresponds to a series of descriptors that



(a) (b) (c)

Fig. 1. The procedure of learning facial anchors by clustering. Red crosses represent keypoints and blue dots represent anchors.

are assigned by Kmeans. Comparing with the scheme of one keypoint with one

descriptor, variance of descriptors coming from the same semantic region of different faces enrich feature representation and are thus less subject to some special face. This property makes unseen face alignment possible. However, noises and even wrong descriptors brought by Kmeans could increase matching risk. Principle Component Analysis (PCA) is effective subspace learning technique and can be used here to represent the intrinsic structure of descriptors. The face model now consists of facial anchors $\{A_i, i = 1, 2, ..., T\}$ and their respective descriptor subspaces $\{S_i, i = 1, 2, ..., T\}$. We can also represent the facial anchor by a set of descriptor exemplars. The different methods of descriptions determine different feature matching criterion, which will be compared in section 4.

3 Partial Face Alignment

3.1 Matching with Shape Constraint

For each frame f of one video, its SIFT feature set consists of M keypoints $\{K_i, i = 1, 2, ..., M\}$ and their respective descriptors $\{D_i, i = 1, 2, ..., M\}$. For each K_i, we compute the distance from the descriptor D_i to each of descriptor subspaces S_j. The facial anchor related to the minimum distance d_1 is the best match and the one related to the second minimum distance d_2 is the second best match. We use the ratio: $r_i = d_1/d_2$ as criterion to estimate the correctness of this match. If $r_i < 0.85$, the match is accepted as a candidate correspondence, otherwise the match is discarded. When this procedure is finished, we sometimes will find that multiple keypoints are matched to a common anchor. In order to guarantee one-to-one correspondence we select the one which has minimum ratio r_i as the candidate correspondence. So we obtain a series of candidate correspondences: $\{C_i : K_i \longleftrightarrow A_j, r_i\}, i = 1, 2, ..., M', M' \leqslant M$.

Even if we adopt the ratio r to threshold matches, there still exist some mismatches due to some uncertain factors such as facial geometric distortion, non-facial features (hair, ear and clothing). We now utilize shape constraint to check candidate correspondences and further kill those outliers.

Given a set of keypoints or anchors, shape constraint is a collection of relative orientations between any two ones and represented as an index matrix. Centered in an anchor A_i, a polar coordinate system can be built so that each of remaining anchors $A_{j,j\neq i}$ gets an angle coordinate. In order to permit somewhat location errors of anchors, angles are quantified into 12 bins by 30^o. Thus each of remaining anchors is labeled a bin index $b_{ij,j\neq i}$ that ranges from 1 to 12. Fig.2 illustrates the procedure. Arranging all these indices are arranged to form a matrix B. After computing the shape constraint matrix B of face model and shape constraint matrix D of novel face, the following algorithm is used to prune the outliers,

1. *Find the C_i which has minimum r_i as the first valid correspondence;*
2. *Set the keypoint K_i of the first valid correspondence as reference point in the novel face;*
3. *For the keypoint K_j of each remaining $C_{j,j\neq i}$;*

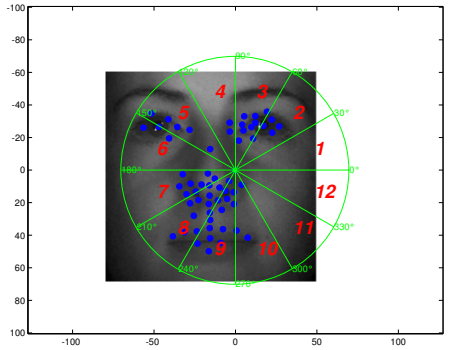
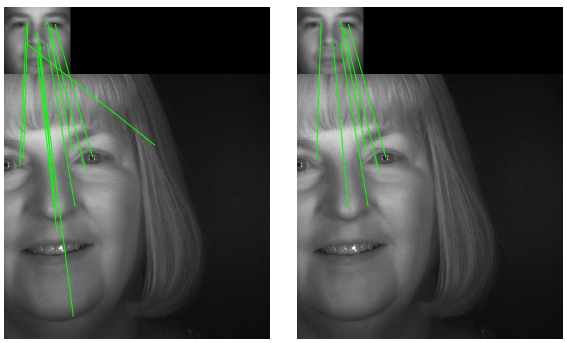


Fig. 2. A diagram for constructing shape constraint in face model. Centered in a given anchor (green dot), a polar coordinate system is plotted in green and orientation indices (red numbers) are assigned to other anchors (blue dots).

4. *By i as row coordinate and j as column coordinate, find index value b_{ij} in B and index value d_{ij} in D;*
5. *If $d_{ij} == b_{ij}$, then accept the current correspondence as the next valid correspondence, else eliminate it;*
6. *Go to the next $C_{j,j\neq i}$.*

An example is given in Fig.3 to illustrate the performance of this algorithm. In



(a) (b)

Fig. 3. Before(a) and After(b) pruning with shape constraint. For each subfigure, the top left small 128×128 image is the template of face model and the bottom image is the novel 512×512 face to be aligned. Correspondences are represented by green line segments which connect the facial anchors and keypoints.

this algorithm, the selection of first valid correspondence determines whether the following selections make sense. If first valid correspondence is a mismatch, all

the following correspondences are then eliminated because of shape constraint. Actually, it only happens to the images that are lack of effective face information, such as the images that only have ears or hairs. The amount of correspondences obtained from one partial face is relatively small. It probably results in unstable similarity solution. Thus frame-by-frame alignment is not the best choice. In the following, we propose to explore the temporal constraint within video to obtain more correspondence inliers. It will improve the alignment performance.

3.2 Alignment with Temporal Constraint

The top row of Fig.4 lists some faces of one sequence. It is clear that their poses varies continuously and lightly. This kind of frame dependence or temporal constraint can help align the sequence. Consider that each of face has a certain amount of valid correspondences when matching to the face model. If we can make all these correspondences contribute to one uniform similarity transform, inliers will be greatly enriched. To achieve this, the first step is to ensure faces within a video have the same pose. Given a video sequence, $\{f_1, f_2, ..., f_N\}$, we extract SIFT features of each frame and then select the frame with most SIFT features as the reference frame f_r so that we can save different kinds of facial features as many as possible. For each of remaining frames f_i within sequence, we establish SIFT feature correspondences between f_i and the reference frame f_r. If the number of correspondences is less than two, then f_i is discarded; otherwise a similarity transform T_i can be solved with four parameters: scaling factor s_i, rotation angle θ_i, x translation t_{xi} and y translation t_{yi}. As a result, we obtain a series of new faces,

$$f_i' = T_i(f_i, \Theta_i), i = 1, 2, ..., N', N' \leqslant N, \tag{1}$$

where f_i' denotes new face and Θ_i denotes four similarity transform parameters: $\{s_i, \theta_i, t_{xi}, t_{yi}\}$. This matching procedure, called 'self-alignment', performs expectably well because of pose continuity and identity consistency. And shape constraint can also help improve the performance. The bottom row of Fig.4 are self-aligned faces with respect to the top row.

Each of self-aligned faces $f_i', i = 1, 2, ..., N'$ within a video can be matched to the face model and contributes a set of M_i valid correspondences $\{C_j^i\}, j = 1, 2, ..., M_i$ by using the method in section 3.1. From all of valid correspondences $\{C_j{}^i\}, j = 1, 2, ..., M_i, i = 1, 2, ..., N'$, an uniform similarity transform T_0 can be further solved by RANSAC algorithm. Final aligned faces are

$$f_i'' = T_0(f_i', \Theta_0) = T_0(T_i(f_i, \Theta_i), \Theta_0), i = 1, 2, ..., N', \tag{2}$$

where Θ_0 denotes four parameters of uniform similarity transform : scaling factor s_0, rotation angle θ_0, x translation t_{x0} and y translation t_{y0}.

4 Experiments

In this section, we conduct experiments on MBGC NIR video sequences to evaluate our method's performance.

(a) Some sample images from one video

(b) Aligned images

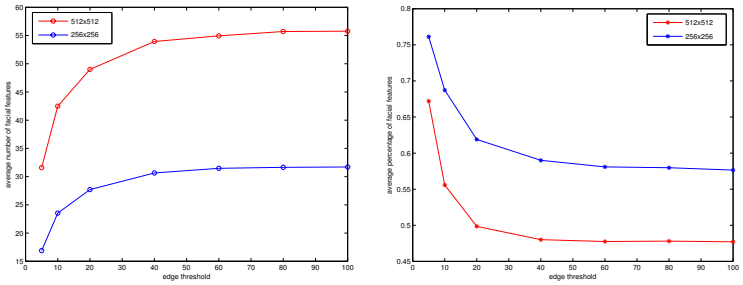
Fig. 4. Alignment within a video. The top row are some original sample images from one video with small pose variation and the bottom row are respective aligned images (in blue box) referred to the last image (in red box) which has most SIFT features in this video.

MBGC sponsored by multiple U.S. Government Agencies seeks to investigate, test and improve performance of face and iris recognition technology on both still and video imagery through a series of challenge problems. One of these problem is to recognize people from near infrared video sequences. There are together 139 sequences which consist of 2286 images with resolution 2048×2048. Sequences are acquired as people walk through a portal and consequently faces in sequences are partial, skewed and even missed. We select 249 holistic frontal faces as training set to learn face model and the remaining 2037 images are used as test set.

For the sake of computational efficiency, all the images are firstly compressed to the resolution 512×512 and 256×256, respectively. An important parameter that affects SIFT features is 'edge threshold'. Edges are poorly defined in different-of-Gaussian function and are less stable to be candidate keypoints. The bigger the 'edge threshold' value is, the more edge-like SIFT features are accepted. As a result, more non-facial features are included, such as hair, ears, collar. In Fig.5(a) and Fig.5(b), we see that when edge threshold increases the average numbers of facial features accordingly increase but rather the average percentages of facial features decreases. When edge threshold passes 20 the average number and percentage tend to be stable. Thus, we set edge threshold to 20 in the following experiments.

In section 2, two kinds of description of facial anchor are introduced. These two methods result in different matching and further alignment performance. As four parameters should be solved, an alignment needs at less 2 valid correspondences. Therefore, if an image has more than 2 valid correspondences, the face is 'detected' and process to be aligned. In order to evaluate alignment performance, we define average square keypoint displacement d_i:

$$d_i = \sqrt{\frac{1}{M_i} \sum_{j=1}^{M_i} \|K'_j - A_j\|^2}, \tag{3}$$



(a) The average numbers of facial features vs. edge threshold.

(b) The average percentages of facial features in total features vs. edge threshold.

Fig. 5. Edge threshold influence on facial SIFT features.

where K'_j is aligned keypoint, A_j is its corresponding anchor and M_i is the number of correspondences. We use R, which has been defined as the smallest displacement tolerance in section 2, as threshold. If $d_i \leq R$, then the alignment is 'correct'; or else the alignment is 'incorrect'.

Table 1. Comparisons of different criterion in detection rate, correct rate and the number of incorrect alignments. Best results are red-marked.

	Resolution	Detection rate	Correct rate	Incorrect alignments
NN	512×512	49.3%	88.6%	129
	256×256	41.5%	95.7%	41
PCA	512×512	45.8%	98.3%	18
	256×256	38.1%	98.1%	17

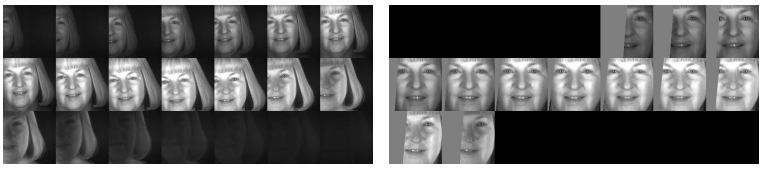
Tab.1 lists frame-by-frame alignment results under NN and PCA criteria. The best correct rate 98.3% is achieved under PCA criterion and the best detection rate 49.3% is achieved under NN criterion. Note that many images in test video database are lack of effective facial parts, like first and last several images in Fig.6(a). Thus, the detection rate is relatively low. NN seems to be more sensitive to the spatial resolution as the correct rate increase from 88.6% to 95.7% by increasing the spatial resolution from 512×512 to 256×256. PCA is less affected by this parameter as the correct rates in 512×512 and 256×256 are similar.

We now compare the performance of frame-by-frame alignment and alignment with temporal constraint in Tab.2. The spatial resolution is set to 512×512 in this experiment. In the first step of alignment with temporal constraint, called self-alignment, each frame is more easily matched to the reference frame than to an uniform face model and thus more facial images are correctly self-aligned and processed to the next step, called joint-alignment. In this step all the de-

Table 2. Comparisons of alignment with and without temporal constraint in detection rate, correct rate and the number of incorrect alignments. Best results are red-marked.

	Temporal constraint	Detection rate	Correct rate	Incorrect alignments
NN	No	49.3%	88.6%	129
	Yes	49.4%	97.7%	26
PCA	No	45.8%	98.3%	18
	Yes	47.9%	99.9%	8

tected faces are matched to the face model and an uniform similarity transform is solved to align all self-aligned faces. Both of best detection rate and correct rate are achieved with temporal constraint. Note that the error in alignment with temporal constraint consists of two part. The one is introduced in the step of self-alignment and the other is generated in the step of joint alignment. For example, under PCA criterion and with temporal constraint, there are 8 incorrect alignments, 6 of which are self-alignment errors and 2 of which are coming from one video and are introduced in joint alignment. Finally, we show aligned images in the video plotted in Fig.6(a) by joint alignment in Fig.6(b).



(a) Original video sequence (b) Aligned video sequence

Fig. 6. An example of video sequence alignment. Original video sequence has 21 frames. After alignment, 12 frontal partial faces are obtained and 9 images have been discarded which are illustrated by black ones.

5 Conclusions

In this paper, we introduce a novel problem of aligning partial faces in NIR video sequences and propose an effective solution. Our method has three novelties. First, an uniform face model is learned from a set of training faces by clustering analysis, which makes unseen face alignment possible. Second, shape constraint is used to eliminate outliers while matching the frames to model. Third, guaranteed by the temporal constraint we develop an scheme of joint alignment. Its results are shown to perform well. In the future, we will concentrate in decreasing the time complexity and develop more robust solution of similarity transform than RANSAC.

Acknowledgements

This work was supported by the following fundings: National Natural Science Foundation Project #60518002, National Science and Technology Support Program Project #2006BAK08B06, National Hi-Tech (863) Program Projects #2006AA01Z192, #2006AA01Z193, and #2008AA01Z124, Chinese Academy of Sciences 100 People Project, and AuthenMetric R&D Funds. We also would like to thank Andrea Vedaldi for his open source code of SIFT implementation.

References

1. Cootes, T., Cooper, D., Taylor, C., Gramham, J.: "Active shape models - their training and application". Computer Vision and Image Understanding **61** (1995) 38–59
2. Cootes, T., Adwards, G., Taylor, C.: "Active appearance model". IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(6) (June 2001) 681–685
3. Lowe, D.G.: "Distinctive image features from scale-invariant keypoints". International Journal of Computer Vision **60**(2) (2004)
4. Slot, K., Kim, H.: "Keypoints derivation for object class detection with sift algorithm". In: Proceedings of Artificial Intelligence and Soft Computing. Volume 4029. (2006) 850–859
5. Carneiro, G., Jepson, A.D.: "Pruning local feature correspondences using shape context". In: Proceedings of IEEE International Conference on Image Processing. (2004)
6. Belongie, S., Malik, J., Puzicha, J.: "Shape matching and object recognition using shape contexts". IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(4) (April 2002) 509–522
7. Lowe, D.G.: "Local feature view clustering for 3d object recognition". In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2001) 682–688
8. NIST: Multiple Biometric Grand Challenge (MBGC). http://face.nist.gov/mbgc (2008)
9. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: "A comparison of affine region detectors". International Journal of Computer Vision **65**(12) (December 2005) 43–72
10. Mikolajczyk, K., Schmid, C.: "A performance evaluation of local descriptors". IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(10) (October 2005) 1615–1630