Anchored Deformable Face Ensemble Alignment

Xin Cheng[1], Sridha Sridharan[1], Jason Saraghi[2], and Simon Lucey[1,2]

Queensland University of Technology, Australia[1]
The Commonwealth Scientific and Industrial Research Organisation[2]
{x2.cheng,s.sridharan}@qut.edu.au,
{jason.saraghi,simon.lucey}@csiro.au

Abstract. At present, many approaches have been proposed for deformable face alignment with varying degrees of success. However, the common drawback to nearly all these approaches is the inaccurate landmark registrations. The registration errors which occur are predominantly heterogeneous (i.e. low error for some frames in a sequence and higher error for others). In this paper we propose an approach for simultaneously aligning an ensemble of deformable face images stemming from the same subject given noisy heterogeneous landmark estimates. We propose that these initial noisy landmark estimates can be used as an "anchor" in conjunction with known state-of-the-art objectives for unsupervised image ensemble alignment. Impressive alignment performance is obtained using well known deformable face fitting algorithms as "anchors".

1 Introduction

Alignment of deformable faces in an image/video has attracted great interest in the computer vision community motivated by its wide range of applications, such as face recognition, facial expression analysis, facial animation, and audio-visual speech recognition. It is a difficult problem as it involves an optimization in high dimensions where appearance can vary greatly between instances of the object due to lighting conditions, facial hair, pose, age, ethnicity, image noise, and resolution. Many approaches have been proposed for this problem with varying degrees of success. Popular models include Active Appearance Models (AAMs) [1], Active Shape Models (ASMs) [2] and Constrained Local Models (CLMs) [3].

Of particular interest in this paper is the task of performing deformable face fitting across an ensemble of facial images stemming from the same subject. This ensemble of images is not necessary causal, so the facial images can be taken from non-uniform samples in time. Appearance consistency between images in the ensemble is an obvious cue/constraint for this problem. We refer to appearance consistency here as the concept that all faces in an image ensemble are of similar appearance given that they are registered to the same coordinate frame of reference. Employing appearance consistency blindly, however, can lead to poor performance for two reasons. First, an ensemble of face images can be considered aligned to a similar geometric frame of reference without looking like a face

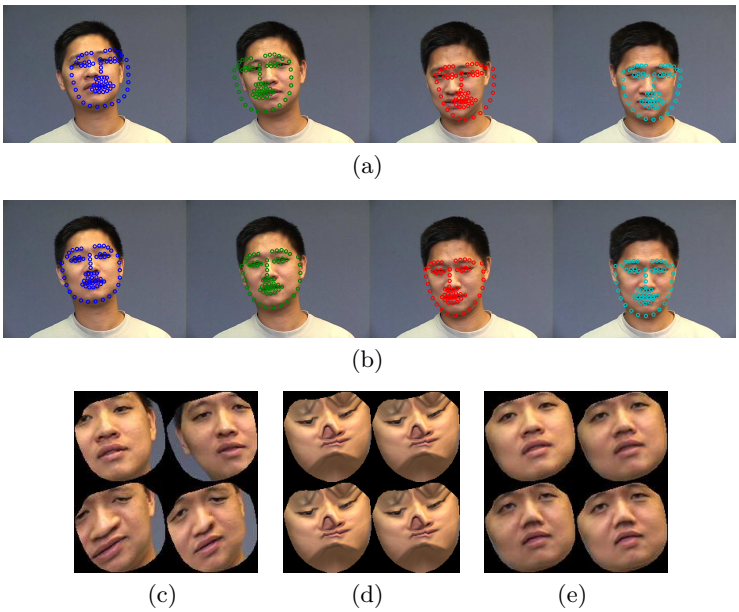


Fig. 1. (a) 4 (out of 40) IJAGS images with very noisy initial alignment. (b) images aligned by the proposed method. (c) faces transformed from the noisy initialization to a reference shape frame. (d) drift (faces aligned without anchoring). (e) faces transformed from the aligned registrations to the reference shape frame.

(see Figure 1(d)), as their is nothing "anchoring" the relative alignment. Second, even though the identity across facial images is constant, other factors are not; including pose, illumination, disappearance/appearance of pixels (e.g. oral cavity opening, eye blinks, occlusions). Due to these problems, most deformable face fitting approaches [1–3] assume appearance independence between frames, instead relying on models / templates learned from offline labelled face datasets. Although providing good performance in general, these approaches often yield imperfect/noisy estimates of landmark positions.

The problem of deformable face fitting across an ensemble of facial images is closely related to the problem of unsupervised image ensemble alignment [4–6]. Recently, an approach referred to as Robust Alignment by Sparse and Low-rank (RASL) decomposition was proposed by Peng et al. [6]. RASL has become of increasing interest to vision researchers as it: (i) can robustly handle variations in illumination through a rank minimization strategy, and (ii) can model outliers and occlusions using an $\mathcal{L}1$ error term. However, RASL cannot manage deformable face fitting in its current framework. In this paper we make three central contributions. First, we introduce an efficient compositional piece-wise affine framework to RASL so as to handle the deformable face fitting task. Second, we propose that noisy estimates from a canonical face fitting algorithm (e.g. AAM, ASM, CLM, etc.) can be introduced into the RASL objective as

an "anchoring" term to remove the improper face warping. Third, we demonstrate state of the art performance for deformable face fitting on the IJAGS face datasets (see Figure 1).

2 RASL

RASL is a specific application of an earlier work called Robust Principal Component Analysis [7]. The authors assume the aligned image ensemble $\mathbf{D} \circ \boldsymbol{\tau}$ is formed by sum of the low rank components $\mathbf{A}$ and sparse errors $\mathbf{E}$,

$$\begin{aligned} \arg\min_{\mathbf{A},\mathbf{E}} & \operatorname{rank}(\mathbf{A}) + \lambda ||\mathbf{E}||_0 \\ s.t. \quad & \mathbf{D} \circ \boldsymbol{\tau} = \mathbf{A} + \mathbf{E}, \end{aligned} \tag{1}$$

where the image ensemble $\mathbf{D}$ is a matrix where each column is a linearized image, the aligned image ensemble is formed by $\mathbf{D} \circ \boldsymbol{\tau} = [vec(I_1 \circ \boldsymbol{\tau}_1) \cdots vec(I_F \circ \boldsymbol{\tau}_F)]$, in which each $I_i \circ \boldsymbol{\tau}_i$ is image I_i warped by the global transformation $\boldsymbol{\tau}_i$ (e.g. similarity, affine and projective transformation). Since both $\operatorname{rank}(\cdot)$ and $||\cdot||_0$ are non-convex and discontinuous functions, the authors relaxed the convexity by replacing $\operatorname{rank}(\cdot)$ with nuclear norm $||\cdot||_*$ and $||\cdot||_0$ with $||\cdot||_1$. The transformation parameter $\boldsymbol{\tau}$ is optimized by an additive framework,

$$\begin{aligned} & \arg\min_{\mathbf{A},\mathbf{E},\Delta\boldsymbol{\tau}} ||\mathbf{A}||_* + \lambda ||\mathbf{E}||_1 \\ s.t. \quad & \mathbf{D} \circ (\boldsymbol{\tau} + \Delta\boldsymbol{\tau}) = \mathbf{D} \circ \boldsymbol{\tau} + \mathbf{J}\Delta\boldsymbol{\tau} = \mathbf{A} + \mathbf{E}, \end{aligned} \tag{2}$$

where $\mathbf{J}$ is the image Jacobian [8] evaluated at the current transformation $\boldsymbol{\tau}$, $\Delta\boldsymbol{\tau}$ is the additive transformation parameter. In every iteration, the parameters are updated as $\boldsymbol{\tau} = \boldsymbol{\tau} + \Delta\boldsymbol{\tau}$. The conventional RASL method is limited to only global transformations. It is not suitable for face alignment tasks as the global transformations lose the geometric information when applied to non-planar object (i.e. human face). Furthermore, in RASL, the Jacobian matrix $\mathbf{J}$ is evaluated at the updated transformation parameter $\boldsymbol{\tau}$ iteratively. This incurs significant cost in computation time, especially for an ensemble with a large number of images.

3 Anchored Deformable Face Alignment

In this Section, we introduce our deformable face ensemble alignment method. We firstly extend RASL by adding a compositional piece-wise-affine transformation function. We then introduce a landmark anchoring penalty to prevent landmarks drift (as shown in Figure 1(d)) after convergence.

3.1 Compositional Alignment

The shape of a deformable subject can be modelled by a mesh, more specifically, by the landmark locations. Mathematically, we define the shape $\mathbf{s}$ with a mesh with v vertices,

$$\mathbf{s} = (x_1, y_1, x_2, y_2, \cdots, x_v, y_v)^{\mathrm{T}}. \tag{3}$$

By applying PCA to a hand labelled face dataset, the shapes of the face can be interpreted by a number of shape parameters $\mathbf{p} = [p_1, p_2, \cdots p_n]^{\mathrm{T}}$, then

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i. \tag{4}$$

Each shape $\mathbf{s}$ contains a large number of triangles defined by vertices. Each pair of corresponding triangles from two shapes define a unique affine transformation. To warp a pixel $\mathbf{x}$, we firstly identify which triangle $\mathbf{x}$ belongs to, then we warp it with the affine transformation of that particular triangle. This method is referred to as piece-wise affine transformation. The conventional RASL exploits an additive framework, in which, the Jacobian of the transformation function $\frac{\partial}{\partial \boldsymbol{\tau}} \mathcal{W}(\boldsymbol{\tau})$ is evaluated at $\boldsymbol{\tau}$. In cases of global transformations as in [6], the Jacobian is constant at all parameters $\boldsymbol{\tau}$. However, for more complicated transformations such as piece-wise affine transformation, the transformation is non-linear, the Jacobian has to be recomputed in every iteration as $\mathbf{p}$ is updated iteratively. This will result in a significant computational cost. The compositional framework provides an alternative to the additive methods. Rather than updating the transformation $\boldsymbol{\tau}$ by $\boldsymbol{\tau} + \Delta\boldsymbol{\tau}$, it updates the transformed images $\mathbf{D} \circ \mathbf{p}$ by $\mathbf{D} \circ \mathbf{p} \circ \Delta\mathbf{p}$. In this framework, the objective function Eqn. **2** can be rewritten as,

$$\begin{aligned} &\arg \min_{\mathbf{A},\mathbf{E},\Delta\mathbf{p}} ||\mathbf{A}||_* + \lambda||\mathbf{E}||_1 \\ &s.t. \quad \mathbf{D} \circ \mathbf{p} \circ \Delta\mathbf{p} = \mathbf{D} \circ \mathbf{p} + \mathbf{J}\Delta\mathbf{p} = \mathbf{A} + \mathbf{E}. \end{aligned} \tag{5}$$

The image Jacobian matrix $\mathbf{J}$ is formed as,

$$\mathbf{J} = \nabla I(\mathbf{p}) \frac{\partial \mathcal{W}(\mathbf{0})}{\partial \mathbf{p}}, \tag{6}$$

where $\nabla I(\mathbf{p})$ is the image gradient evaluated at $\mathbf{p}$. This gradient has to be recalculated every iteration, however, it is an efficient process compared with recomputing the Jacobian of the piece-wise-affine transformation, $\frac{\partial}{\partial \mathbf{p}} \mathcal{W}$. Fortunately in compositional alignment, since the Jacobian of transformation function is always evaluated at $\mathbf{0}$, it can be precomputed as it only needs to be computed once.

3.2 Anchored RASL

Since there is no prior knowledge of facial appearance exploited, without anchoring, the process will deform the subject's face arbitrarily to find the minimum rank, in nearly all instances resulting in a false alignment. In the proposed method, we introduce a vertex anchoring method using the $\mathcal{L}2$-norm, whose objective function is,

$$\begin{aligned} \arg \min_{\mathbf{A},\mathbf{E},\Delta\mathbf{p}} &||\mathbf{A}||_* + \lambda_1||\mathbf{E}||_1 + \lambda_2||\mathbf{X} + \boldsymbol{\Phi}\Delta\mathbf{p} - \mathbf{S}||_2^2 \\ &s.t. \quad \mathbf{D} \circ \mathbf{p} \circ \Delta\mathbf{p} = \mathbf{D} \circ \mathbf{p} + \mathbf{J}\Delta\mathbf{p} = \mathbf{A} + \mathbf{E}, \end{aligned} \tag{7}$$

where $\mathbf{X}$ is the locations of the current vertices, $\boldsymbol{\Phi}$ is the shape basis matrix (each column in $\boldsymbol{\Phi}$ is a eigenvector of shape), and $\mathbf{S}$ is the anchoring points. In this work we use the initial alignment as anchoring points to avoid the need for additional knowledge. Our experiment shows that although the anchoring points are noisy in terms of landmark locations, they are still able to stabilize the process by stopping alignment from drifting. Our objective function Eqn. **7** can be optimized efficiently by the Augmented Lagrangian Method [6],

$$\mathcal{L}(\mathbf{A}, \mathbf{E}, \Delta\mathbf{p}, \mathbf{Y}) = ||\mathbf{A}||_* + \lambda_1||\mathbf{E}||_1 + \lambda_2||\mathbf{X} + \boldsymbol{\Phi}\Delta\mathbf{p} - \mathbf{S}||_2^2 + < \mathbf{Y}, \mathbf{D} \circ \mathbf{p} + \mathbf{J}\Delta\mathbf{p} - \mathbf{A} - \mathbf{E} > + \frac{\mu}{2}||\mathbf{D} \circ \mathbf{p} + \mathbf{J}\Delta\mathbf{p} - \mathbf{A} - \mathbf{E}||_2^2, \quad (8)$$

where $\mathbf{Y}$ is the Lagrangian Multiplier, μ is a positive scaler, $< \cdot, \cdot >$ is matrix inner product. Then in every iteration, the new values of $\mathbf{A}$, $\mathbf{E}$, $\Delta\mathbf{p}$ and $\mathbf{Y}$ can be determined by alternating,

$$\mathbf{A}^{k+1} = \arg\min_{\mathbf{A}} \mathcal{L}(\mathbf{A}, \mathbf{E}^k, \Delta\mathbf{p}^k, \mathbf{Y}^k) \quad (9)$$

$$\mathbf{E}^{k+1} = \arg\min_{\mathbf{E}} \mathcal{L}(\mathbf{A}^{k+1}, \mathbf{E}, \Delta\mathbf{p}^k, \mathbf{Y}^k) \quad (10)$$

$$\Delta\mathbf{p}^{k+1} = \arg\min_{\Delta\mathbf{p}} \mathcal{L}(\mathbf{A}^{k+1}, \mathbf{E}^{k+1}, \Delta\mathbf{p}, \mathbf{Y}^k) \quad (11)$$

$$\mathbf{Y}^{k+1} = \mathbf{Y}^k + \mu(\mathbf{D} \circ \mathbf{p} + \mathbf{J}\Delta\mathbf{p}^{k+1} - \mathbf{A}^{k+1} - \mathbf{E}^{k+1}). \quad (12)$$

The $\mathbf{A}^{k+1}$ and $\mathbf{E}^{k+1}$ can be determined using the soft threshold method as described in [7], The update of parameters $\Delta\mathbf{p}$ can be found by,

$$\frac{\partial}{\partial\Delta\mathbf{p}}\mathcal{L}(\mathbf{A}^{k+1}, \mathbf{E}^{k+1}, \Delta\mathbf{p}, \mathbf{Y}) = \frac{\partial}{\partial\Delta\mathbf{p}}(\lambda_2||\mathbf{X} + \boldsymbol{\Phi}\Delta\mathbf{p} - \mathbf{S}||_2^2 + \frac{\mu}{2}||\mathbf{D} \circ \mathbf{p} + \mathbf{J}\Delta\mathbf{p} - \mathbf{A}^{k+1} - \mathbf{E}^{k+1} + \frac{1}{\mu}\mathbf{Y}^k||_2^2) = 0, \quad (13)$$

then we have,

$$\Delta\mathbf{p}^{k+1} = (2\lambda_2\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \mu\mathbf{J}^{\mathrm{T}}\mathbf{J})^{-1}[2\lambda_2\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{S} - \mathbf{X}) + \mu\mathbf{J}^{\mathrm{T}}(\mathbf{A}^{k+1} + \mathbf{E}^{k+1} - \frac{1}{\mu}\mathbf{Y}^k - \mathbf{D} \circ \mathbf{p})]. \quad (14)$$

The overall algorithm is described in Algorithm 1.

4 Experiments

In this section, we evaluate the performance of our Anchored RASL method on a variety of face alignment tasks. The face shape model employed in the evaluation was obtained by a training process from all subjects of the IJAGS database and MultiPIE [9] database (5 subjects of IJAGS and 346 subjects in MultiPIE, with varying head poses and facial expressions). The shape model consists of 19 degrees of freedom with 66 landmark points. The image in the

Algorithm 1 Face refinement using Anchored RASL

Input: the initial landmarks $\mathbf{S}$, weights λ_1, λ_2, shape basis $\boldsymbol{\Phi}$, total number of frames F, each frame has P points.
Solve for the initial shape parameter, $\mathbf{p} = \text{eval}(\mathbf{S},\boldsymbol{\Phi})$,
Determine warp Jacobian $\frac{\partial \mathcal{W}(\mathbf{0})}{\partial \mathbf{p}}$.
while not converged **do**
  for $i = 1$ to F **do**
    Warp image, $\widehat{I}_i = I_i \circ p_i$,
    Determine gradient, $\nabla \widehat{I}_i = \text{gradient}(\widehat{I}_i)$,
    Determine Jacobian, $J_i = \nabla \widehat{I}_i \frac{\partial \mathcal{W}(\mathbf{0})}{\partial \mathbf{p}}$,
    Determine ensemble, $\mathbf{Dp}(i,:) = \text{vec}(\widehat{I}_i)'$,
    Determine the current mesh, $\mathbf{X} = \boldsymbol{\Phi}\mathbf{p}$.
  end for
  Solve for $\Delta\mathbf{p}$ using

$$\arg \min_{\mathbf{A},\mathbf{E},\Delta\mathbf{p}} ||\mathbf{A}||_* + \lambda_1||\mathbf{E}||_1 + \lambda_2||\mathbf{X} + \boldsymbol{\Phi}\Delta\mathbf{p} - \mathbf{S}||_2^2$$
$$s.t. \qquad \mathbf{Dp} + \mathbf{J}\Delta\mathbf{p} = \mathbf{A} + \mathbf{E}, \tag{15}$$

  Update shape parameter $\mathbf{p} = \mathbf{p} \circ \Delta\mathbf{p}$.
end while

reference shape frame was scaled to 10,000 RGB pixels. The weight, λ_1 was selected using the same strategy as in [6], $\lambda_1 = 1/\sqrt{m}$, where m is the number of pixels in each aligned image (30,000 in our case). The experiment result shows that the best performance was found when using $\lambda_2 = 0.03/\sqrt{n}$, where n is the number of landmark points in every frame (66 in our implementation). The CLMs tracker we employed in the experiment was implemented by [10]. The shape model and the local features of the CLMs tracker were trained with all subjects of MultiPIE database [9].

4.1 Anchored RASL Vs. Unanchored RASL

To validate the importance of the anchoring term, we evaluated the performance of our anchored RASL method and the conventional RASL on image sequences with synthetic noisy landmark registrations. 40 frames of a single subject with large head pose variations were selected from the IJAGS database. We randomly selected a subset of $n = 32$ frames (equivalent to 80% of the frames), and perturb the annotated ground truth landmarks with synthetic errors. For each selected frame, a random synthetic error $E_i \in \mathcal{N}(0, \sigma^2)$ was added to all landmark points to produce a global alignment offset. In the experiment, we generated test cases with different geometric errors in the anchor points by increasing the standard deviation σ. The performance of our Anchored RASL method and the conventional unanchored RASL method were compared with the RMS geometric errors (shown in Figure 2(a)) and the nuclear norms (shown in Figure 2(b)). The experimental results show that the conventional RASL searched the minimum

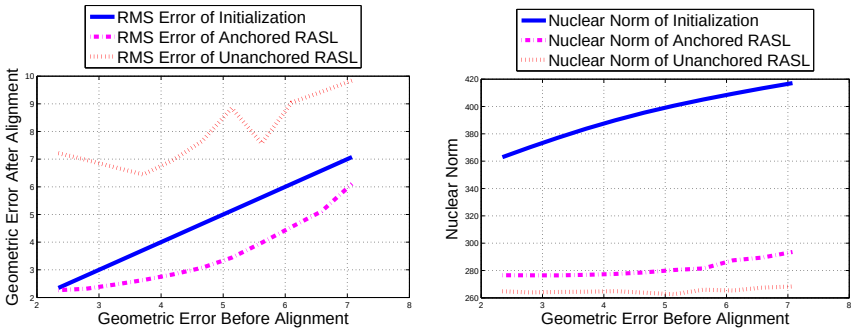


Fig. 2. (a) The RMS geometric errors; (b) The nuclear norms. The conventional RASL is not suitable for deformable face alignment as it searches for the lowest nuclear norm by blindly distorting the faces. To address this problem our anchored approach constrains the landmarks in certain regions to ensure a good alignment.

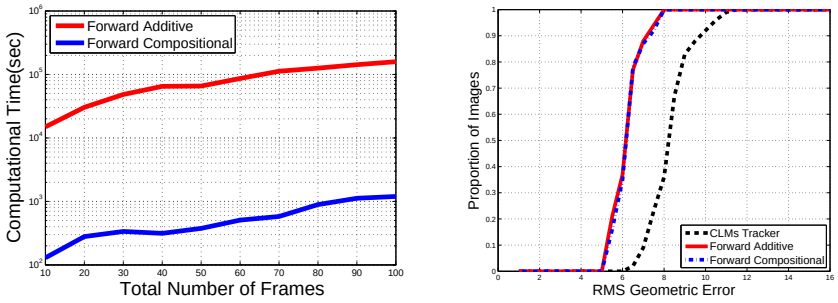
nuclear norm by arbitrarily distorting the faces in each frame. Our anchored RASL method is able to maintain the landmark points in reasonable locations to stop the improper distortions, in order to ensure a good alignment.

4.2 The Efficiency Evaluation

To verify the efficiency improvement of our Compositional Anchored RASL method from the conventional Additive method [6], we compared the computational time and the fitting performance of each method with a sequence of 100 IJAGS face images. The alignment was initialized and anchored by landmark points determined by the state-of-the-art CLMs tracker [3, 10]. The computational time for aligning different number of frames were tracked and presented in Figure 3(a). The fitting performance of the two methods were demonstrated in Figure 3(b). The experimental results show that both the additive method and the compositional method are able to refine the alignment from the state-of-the-art CLMs tracker. The proposed compositional method is able to reduce approximately 99% of the computational cost of the conventional additive method, while maintaining identical fitting performance.

4.3 Visualization

In order to visually inspect the effectiveness of the proposed method, we have selected two simulation results for visualization. The first simulation is conducted using IJAGS database, 40 frames were selected using the same criterion as in the previous section. The σ of the simulated error as defined in Section 4.1 is set to approximately 5% of the average face size. The normalized faces (face transformed from the original image shape frame to the reference shape frame) of the initial alignment and the refined alignment are present in Figure 4(a) and



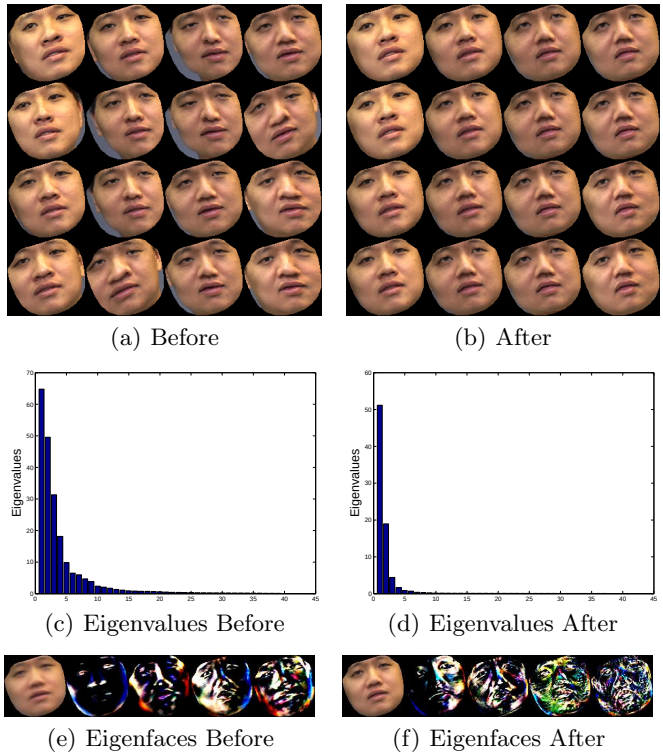
(a) (b)

Fig. 3. (a) The computational time with different number of frames; (b) The fitting performance of the two methods when processing 100 frames. It can be observed that the compositional method can reduce the computational cost significantly while maintaining identical fitting performance.

Figure 4(b) respectively. The eigenvalues of the presented normalized faces were determined by principal component analysis and demonstrated in Figure 4(c) and Figure 4(d). The mean face and the first 4 eigenfaces are presented in Figure 4(e) and Figure 4(f). In order to evaluate our approach with low image quality, the second simulation was conducted using 40 frames from the MultiPIE database selected with strong illumination variations. The sequence was converted to grayscale and blurred by Gaussian kernel to lower the image quality. The sequence was initialized by Gaussian errors with σ set to 5% of the average face size. The same set of visualizations are presented in Figure 5.It can be observed that by using our Anchored RASL alignment method, the initial coarse alignments of both dataset are refined. The eigenvalues of the refined sequences are narrowly distributed to the first few Eigenspaces. The mean faces of the refined sequences are very clear whereas the mean faces of the initial alignment are very blurred. The Eigenfaces of the refined sequence is more random. This is because there are fewer appearance variations of the well aligned faces than the misaligned faces.

5 Conclusion

In this paper, we introduced a new anchored method for deformable image ensemble alignment. This method introduced an efficient compositional piece-wise affine framework to RASL that extends the benefits of RASL to deformable face fitting. This includes robustness to illumination variation through rank minimization and ability to model outliers and occlusions using an $\mathcal{L}1$-norm term. We evaluated our method using a subset of IJAGS database with pose variations and a subset of MultiPIE database with strong illumination variations. Impressive experimental results were demonstrated with different image condi-



(a) Before (b) After

(c) Eigenvalues Before (d) Eigenvalues After

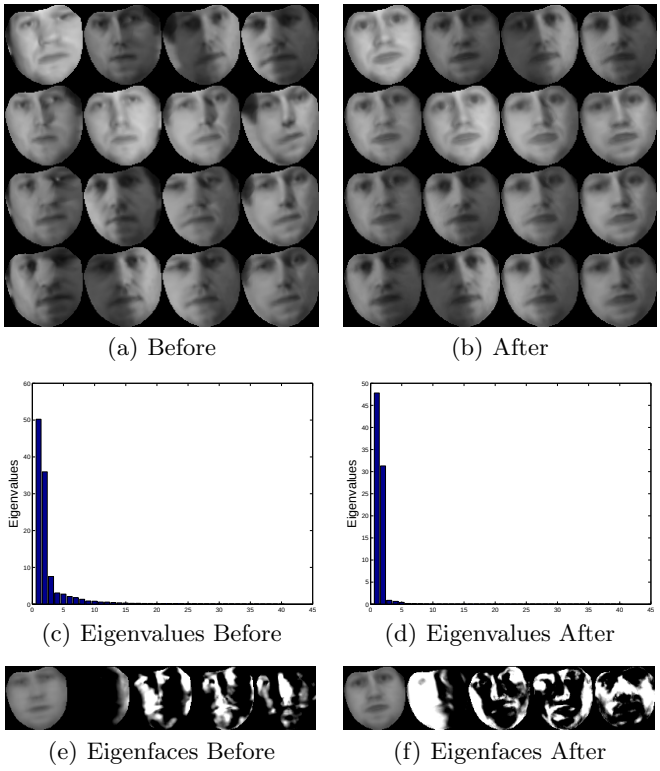
(e) Eigenfaces Before (f) Eigenfaces After

Fig. 4. A sequence of 40 frames are selected from the IJAGS database. 80% of the frames were perturbed by the Gaussian errors with σ set to approximately 5% of the average face size for initialization and anchoring.

tions. The anchoring method demonstrated strong ability to nonrigidly align an ensemble of face images without improper distortion of facial appearance.

References

1. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60** (2004) 135 – 164
2. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models their training and application. Comput. Vis. Image Underst. **61** (1995) 38–59
3. Saragih, J.M., Lucey, S., Cohn, J.: Face alignment through subspace constrained mean-shifts. In: International Conference of Computer Vision (ICCV). (2009)
4. Learned-Miller, E.G.: Data driven image models through continuous joint alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006) 236–250
5. Cox, M., Lucey, S., Sridharan, S., Cohn, J.: Least squares congealing for unsupervised alignment of images. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). (2008)



(a) Before (b) After

(c) Eigenvalues Before (d) Eigenvalues After

(e) Eigenfaces Before (f) Eigenfaces After

Fig. 5. A sequence of 40 frames are selected from the MultiPIE database. The sequence was converted to grayscale and blurred by the Gaussian kernel. 80% of the frames were perturbed by the Gaussian errors with σ set to approximately 5% of the average face size for initialization and anchoring.

6. Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. (2010) 763 –770
7. Wright, J., MA, Y., GANESH, A., RAO, S.: Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices via Convex Optimization. Proceedings of Neural Information Processing Systems (NIPS) (2009)
8. Baker, S., Matthews, I.: Lucas-kanade 20 years on: A unifying framework. International Journal of Computer Vision **56** (2004) 221 – 255
9. Gross, R., Matthews, I., Cohn, J.F., Kanade, T., Baker, S.: Multi-PIE. Image and Vision Computing (2009)
10. Saraghi, J.: Facetracker. In: http://web.mac.com/jsaragih/FaceTracker /FaceTracker.html. (2011)