Joint Face Alignment with a Generic Deformable Face Model

Cong Zhao Wai-Kuen Cham Xiaogang Wang
Department of Electronic Engineering
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
{czhao, wkcham, xgwang}@ee.cuhk.edu.hk

Abstract

As having multiple images of an object is practically convenient nowadays, to jointly align them is important for subsequent studies and a wide range of applications. In this paper, we propose a model-based approach to jointly align a batch of images of a face undergoing a variety of geometric and appearance variations. The principal idea is to model the non-rigid deformation of a face by means of a learned deformable model. Different from existing model-based methods such as Active Appearance Models, the proposed one does not rely on an accurate appearance model built from a training set. We propose a robust fitting method that simultaneously identifies the appearance space of the input face and brings the images into alignment. The experiments conducted on images in the wild in comparison with competing methods demonstrate the effectiveness of our method in joint alignment of complex objects like human faces.

1. Introduction

Nowadays, with great reduction in the cost of image acquisition devices and great development in the capacity of data storage, to have more than one images of an object becomes increasingly more practical and convenient. For example, internet photo sharing sites such as Facebook and Flickr, as well as personal photo library software like iPhoto and Picasa, most likely hold multiple images of a customer. For video conferencing, sequential output frames of a camera are in general capturing the same face. In addition, group study in medical image engineering also works with multiple instances of a human organ.

On the other hand, to better understand the object in the images, it is important to align them first. In addition to direct employing pair-wise registration techniques, in recent years, joint alignment [1-7] has raised more and more research interest. The objective is to align the images jointly in order to avoid a biased template selection. However, joint alignment of real world objects like human faces remains a difficult problem due to the following challenges: a) they often undergo both rigid transformations and non-rigid deformations. b) Different faces of different people have dramatically different appearances under a variety of illumination conditions, which poses a steep challenge for their alignment.

In this paper, we propose a joint alignment method based on the Active Appearance Models (AAM). The principal idea is to jointly align the batch of images by employing a generic shape model and a generic appearance model both learned from a training set consisting of a variety of faces. We demonstrate that, while the accuracy of the appearance model is critical for the success of conventional AAMs, the proposed method is able to work with a biased and inaccurate one. Specifically, our method identifies the unbiased appearance space of the input face and simultaneously brings the input images into alignment. The goal is achieved under two important assumptions: First, the actual appearance of the given face is linear and low-dimensional. Second, the person-specific space is close to that of generic human facial appearances.

Our main contributions are: a) we propose an effective model-based method to jointly align facial images under both non-rigid deformation and appearance variation. This is essentially different from existing rigid alignment approaches. b) We propose a robust fitting algorithm, so that a generic appearance model trained from a variety of faces can be fit adaptively and consistently to a new unseen face whose appearance cannot be accurately modeled.

The remainder of this paper is organized as follows: We review the related work in Section 2, and introduce the popular AAM algorithm in Section 3 under the background of multi-image alignment. Section 4 formulates the proposed method and provides an efficient solution. We perform experimentation and compare the obtained results with competing methods in Section 5 and finally draw conclusions in Section 6.

2. Related Work

Joint alignment was initiated by the authors in [1] who registered a batch of images in terms of affine transfor-

mations by maximizing the entropy of the image group. Among the succeeding works [2-5], the method proposed in [5] performed much better in handling occlusions and outliers by minimizing the rank of the images and the number of outliers. The above methods succeeded in calculating the rigid transformations among images of an object, however, due to their ignorance of non-rigid deformations, they are restricted from being applied to more ambitious scenarios, for example those in [8-10]. The authors in [11] introduced B-Splines and extended the work [2, 3] to the non-rigid alignment of medical images, however, their method cannot work on objects with significant appearance variations such as human faces.

Among existing non-rigid alignment techniques, model-based ones [13, 14] such as the Active Appearance Model (AAM) are the most popular and effective due to their ability in handling pose, shape and appearance variations. The main idea of AAM is to match a learned deformable model to an input image, so that the natural feature points of the input image are registered with the model, resulting in an alignment. However, one of the most concerned issues for model-based methods is that, their performance greatly reduces when fitting a generic model to an unseen object whose appearance differs significantly from the training samples. This fact restricts AAMs from being applied to objects like human faces under wild conditions. The reason is two-fold: First, the objective function of high dimensional variables has many local minima, which stuck those gradient descents [13, 14] into unwanted solutions. Second, the situation gets worse when the appearance model fails to accurately model the new object.

To investigate this phenomenon, the authors [15] did intensive experimental studies and found that it is much harder for a generic AAM model to either model or be fitted to a new face than a person-specific one does, and the main reason is attributed to the inaccuracy for the generic appearance model in modeling the new object. Many works have been proposed to address this problem, including: the authors in [16] studied different factorization techniques – PCA, ICA, and NMF, and found that they were of quite limited help in performance improvement. The authors in [17] proposed a multi-layer AAM which modeled facial appearances in more detail; however, they did not address fitting it to a new face. In addition, other works were devoted to finding the global optimum of the objective function either by initializing a better start, by constraining the shape between video frames, or by rectifying the gradient during iterations [18]. Furthermore, some works were focused on refining the objective function itself. The authors in [19] proposed to learn a rectified metric from, which explicitly smoothed out local minima and encouraged the global optimum to occur at correct places. The authors in [20] imposed a Gaussian prior term in the objective function of the inverse-compositional algorithm. The term can be determined either by learning from training data, or by solving the system dynamic equation in the context of video tracking.

Different from the above efforts, either on improving the appearance model or on improving the fitting accuracy to one image or frame, in this paper we investigate the problem of joint alignment of a batch of images of an unseen face using a biased generic AAM model. We demonstrate that although the generic appearance model fails to accurately model the new object, making good use of the redundancy among the input images can well compensate for the model error. We propose a method which robustly estimates the appearance of the new object and simultaneously brings the images into alignment. This is in essence different from [15] in which the appearance space is explicitly and supervised learnt from a set of manually labeled training samples of that specific person.

3. Multi-Image Alignment by AAM

To begin with, we first review the algorithm of AAM in the context of multiple image alignment.

3.1. Active Appearance Models

The Active Appearance Model describes an object from its shape and appearance. The shape is denoted as a coordinate vector of a set of landmarks:

$$s = (x_1, y_1, x_2, y_2, ..., x_n, y_n)^T$$

where (x_i, y_i) is the 2D coordinate of the i-th landmark. The appearance is denoted as the set of pixels sampled inside the object region of a reference shape.

For AAMs [13, 14], a common assumption made is that the shape and appearance variations can both be modeled as linear. Therefore, Principal Component Analysis (PCA) can be applied to find the components:

$$s = s_0 + \sum_{i=1}^{n} p_i s_i \quad (1)$$

$$\alpha = \alpha_0 + \sum_{i=1}^{m} \lambda_i \alpha_i \quad (2)$$

where s and α are the shape and the appearance vector in a column-wise manner; s_i and α_i are the i-th Principal Components (PCs) of the shape and appearance model; p_i and λ_i are the corresponding loadings; s_0 and α_0 are the reference shape and appearance vectors which are often regarded as the mean of the training data. Notice that during the training stage, the global transformation and the brightness of the images are normalized beforehand, so that the two linear models are free from global variations.

A template with such shape and appearance variability is known as an AAM model. To fit such a model to an input

image is to minimize the discrepancy between them:

$$\min_{p,q,\lambda} \left\| G(W(I;p);q) - \alpha_0 - A\lambda \right\|_2^2 \qquad (3)$$

where I is the vectorized input image; W is the parameterized warping function which maps I to an appearance vector defined inside the reference shape; G is the global transformation consisting of rotation, translation and scaling; p and q are the unknown parameters of W and G to be determined in the alignment; A is the matrix of PCs describing the principal modes of appearance variations, and λ is the corresponding loading.

Since the objective function is non-linear with respect to both p and q, gradient descent and its variants [13, 14] become the most popular choice. Specifically, they first linearize the problem by first-order approximation, then start searching along descent directions from an initialized position, and then iteratively solve the increments. The difference between them lies in the way of calculating the gradient: numerically or analytically, and the way of composing the increments: forward-additive or inverse-compositional.

3.2. Multi-Images Joint Alignment

We apply the AAM algorithm in multi-image alignment due to its capability in handling both non-rigid deformation and appearance variation. A straightforward idea is to build a generic AAM model from a class of training samples and register it with each input image independently. This leads to an objective function:

$$\min_{p_i,q_i,\lambda_i} \sum_i \tau_i \left\| G(W(I_i;p_i);q_i) - \alpha_0 - A\lambda_i \right\|_2^2 \qquad (4)$$

where τ_i weights the importance of image I_i in the joint alignment. Throughout this paper, we assume equivalent importance and set $\tau_i = 1$ for all i, if no preference is added.

Since the images are considered independently, we can address each sub-problem in (4) separately. We refer to this implementation as the "Independent AAM".

3.3. Limitation of the Generic Appearance Model

It is practically realistic that we wish to build a generic model and apply it in aligning images of any unseen objects. However, it is well-known that a generic model suffers from deficiency in this case and thus "Independent AAM" cannot produce good results.

The reason is that although the shapes of a variety of persons can be well modeled as linear, the appearances of different faces undergoing different illumination are far more complex than linear [15], resulting in violation of the assumptions made in [13, 14]: a) the linear relationship between reconstruction error and model increment, regressed on the training data, does not apply to new data any more. b) The appearance of the new face does not lie in the space learned from the training data, and therefore cannot be "projected out". When these assumptions are violated, gradient descents suffer from biased descending directions, and converge to incorrect solutions. This is one of the reasons that conventional AAMs [13, 14] perform poorly on unseen faces. To illustrate this problem, we show in Fig.3 that the inaccuracy in appearance modeling leads to dramatically different results: (top) using a generic appearance model, (bottom) an unbiased person-specific one.

Furthermore, the inaccuracy of the appearance model can also deteriorate the convexity of the objective function (3). As shown in Fig. 2, the objective function is bumpy and has many local minima around different disturbances of the parameters q (translation, rotation and scaling). This problem becomes even worse in practice when they interact with both shape and appearance parameters.

4. Regularized Fitting for Joint Alignment

While an accurate appearance model is critical for independent AAMs, we demonstrate in this section that an inaccurate model can be well compensated for if we make good use of the images themselves in a joint alignment task.

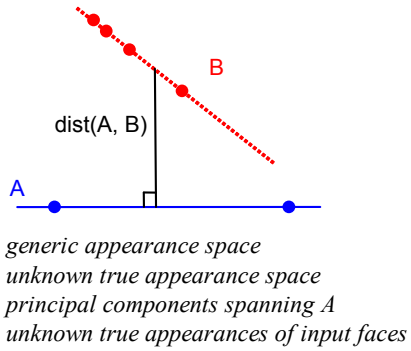
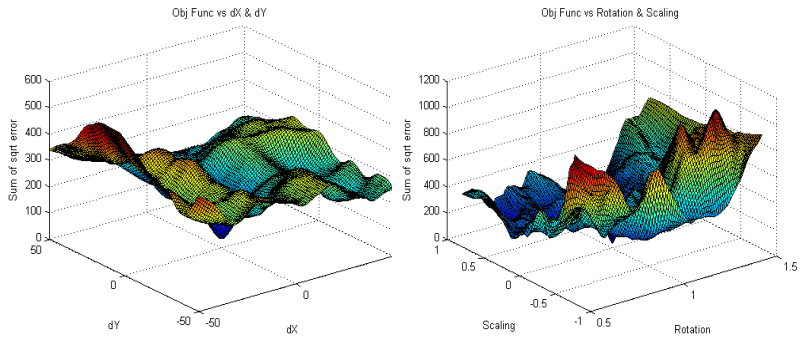


Figure 1: Inaccurate generic appearance space

4.1. Two Important Assumptions

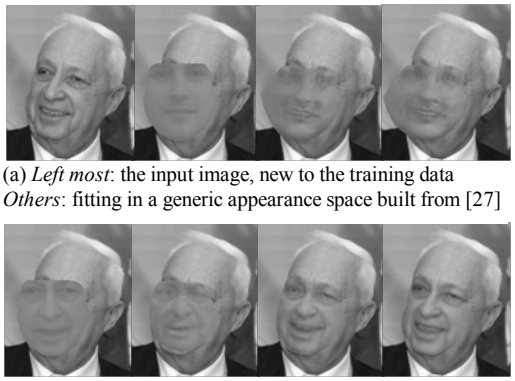
Existing studies [21] have showed that although the generic facial appearances of different faces are much more complex than linear, those of the same person can be well approximated by linear. This fact has been well examined by the success of person-specific AAMs through an intensive experimental study in [15] and its performance in real practice.

This inspires us that, we are guaranteed a good fitting performance if we can build a correct person-specific appearance space for the input face. Motivated by this fact, we make the first important assumption on the person-specific space: *The images of the same face aligned to the reference shape should be linear and low-dimensional*. As mentioned above, this assumption has well examination, and Fig.4



(a) Translation by ± 50 pixels; (b) rotation by ± π/4 and scaling 0.5~1.5 times;

Figure 2: Change of the objective function with disturbance around a ground-truth shape



(a) *Left most*: the input image, new to the training data
Others: fitting in a generic appearance space built from [27]

(b) Fitting in the person-specific appearance space

Figure 3: Comparison of fitting an AAM model in different appearance spaces

(bottom) shows that by using as few as 2 PCs of the specific appearance can we reconstruct the input faces much better than using 98 PCs of the generic appearance.

Furthermore, since both the generic space and the person-specific space span human facial appearances, a reasonable assumption which constitutes our second assumption can be made on the distance between them: *the person-specific space and the generic appearance space should be proximate rather than distant*[1].

Fig.1 gives an illustration to the above two assumptions: The blue line A denotes the generic appearance space learned from the training data of different persons, and the red line B denotes the specific appearance space of the new face to be identified. Intuitively, B coincides with A if the person is included in the training set, and disjoints from A if otherwise. The challenge here, however, is to identify a person-specific B from only a few unaligned images which should "hit" the red dots if well aligned.

Based on the two assumptions, we propose in Section 4 an effective and efficient method which simultaneously finds the person-specific appearance space B and brings the given images into alignment.

4.2. Finding the Appearance Space

The problem of identifying the person-specific space can be formulated as: to find a low-dimensional subspace embedded in the high-dimensional data space that is close to the generic one as much as possible:

$$\min_{p_i,q_i,\lambda_i} \rho\left\{\sum_i \left\|G\left(W\left(I_i;p_i\right);q_i\right)-\alpha_0-A\lambda_i\right\|_2^2\right\} + rank\left(\sum_i G\left(W\left(I_i;p_i\right);q_i\right)e_i^T\right) \quad (5)$$

where the scalar $\rho > 0$ trades off two quantities: one constraints the appearances to be close to generic human faces, and the other encourages them to be linear correlated and group similar. Since the images are jointly considered in the alignment, we refer to (5) as "Joint AAM" in this paper.

It is worth mentioning that, the value of ρ should be determined by the two quantities: if the appearance model is inaccurate leading to larger $dist(A,B)$, then the reliability of the first term should be lower, inducing a smaller ρ; on the other hand, if the images are less correlated or group dissimilar, then the low-rank term should be deemphasized, inducing a larger ρ. We give an empirical way for setting a proper value of ρ in Section 4.6.

4.3. Reformulation

The difficulty in solving (5) lies in the non-convexity and non-continuity of the rank term, which makes minimizing (5) NP-hard. On the other hand, recent theories in Compressive Sensing [22,23] demonstrate a fact that minimizing the rank of a matrix is equivalent, under mild conditions, to minimizing its tightest convex relaxation — the nuclear norm. Therefore, we replace the rank term by its nuclear norm, obtaining an equivalent problem:

$$\min_{x_i,\lambda_i} \rho\left\{\sum_i \left\|T\left(I_i;x_i\right)-\alpha_0-A\lambda_i\right\|_2^2\right\} + \left\|\sum_i T\left(I_i;x_i\right)e_i^T\right\|_* \quad (6)$$

where the nuclear norm of a matrix is defined as the summation of singular values $\|X\|_*=\Sigma_k\sigma_k(X)$. Notice that here we re-represent the combination of the global transformation G and warping function W by T, parameterized by a new variable $x_i = (p_i, q_i)$. And we discuss about this re-representation in Section 4.5.

The problem (6) remains non-linear with respect to variable x_i (stands for p_i and q_i). A common approach to deal with non-linearity is to make first-order approximations, and iteratively solve in (7) the increments Δx_i. We then update the values of x_i by $x_i(k+1) = x_i(k) + \Delta x_i$

[1] We employ the distance from a point set to a space, as the person-specific space is itself spanned by a point set (with unknown warping parameters)

Top row: aligned faces; *second row*: true appearances sampled under the reference shape; *third row*: reconstruction with all PCs, in the biased generic appearance space; *bottom row*: reconstruction with only the first 2 PCs, in the appearance space identified by our method

Figure 4. Reconstruction of appearances in different spaces

$$\min_{\Delta x_i,\lambda_i} \rho\left\{\sum_i \left\|T(I_i;x_i)+J_i\Delta x_i-\alpha_0-A\lambda_i\right\|_2^2\right\} + \left\|\sum_i \left[T(I_i;x_i)+J_i\Delta x_i\right]e_i^T\right\|_* \quad (7)$$

where J_i is the Jacobian matrix of the warped image with respect to the parameters x_i, determined in a similar way as [14].

We further introduce a new variable Y, and reformulate (7) into a semi-definite program:

$$\min_{Y,\Delta x_i,\lambda_i} \rho\left\|Y-A_0-A\Lambda\right\|_F^2+\left\|Y\right\|_* \quad (8)$$
$$\text{s.t. } Y=\sum_i T(I_i;x_i)e_i^T+J_i\Delta x_i e_i^T$$

where A_0 is the matrix composed of replicates of the reference appearance α_0; and Λ is the matrix of appearance coefficient vectors λ_i for all i.

4.4. Efficient Solution

Among the numerical methods, we apply the Augmented Lagragian Method (ALM) which have three appealing advantages: First, it does not require penalty factor σ_k to approach infinity, and is thus free from the problem of ill-conditioning during the growth of σ_k .Second, it converges Q-linearly [26] to the global optimal solution even when the sequence σ_k is bounded. Third, it has only one parameter η, and is easy-tuning. Here we show how the ALM algorithm can be adapted to efficiently solve (8), the augmented Lagragian of which is:

$$\phi(Y,\Lambda,\Delta X,Z,\sigma)=\rho\left\|Y-A_0-A\Lambda\right\|_F^2+\left\|Y\right\|_* + \left\langle Z,g(\Delta X)-Y\right\rangle+\frac{\sigma}{2}\left\|g(\Delta X)-Y\right\|_F^2$$

where $g(\Delta X)=\sum_i T(I_i;x_i)e_i^T+J_i\Delta x_i e_i^T$.

The ALM basically iterates among three update steps – the variable (Y, Λ, ΔX), the Lagragian parameter Z, and the penalty factor σ :

$$\left(Y^{k+1},\Lambda^{k+1},\Delta X^{k+1}\right)=\arg\min_{Y,\Lambda,\Delta X}\phi\left(Y,\Lambda,\Delta X,Z^k\right) \quad (9)$$
$$Z^{k+1}=Z^k+\sigma^k\left[g\left(\Delta X^k\right)-Y^k\right]$$
$$\sigma^{k+1}=\eta\sigma^k$$

The iteration is conducted until the convergence of $g(\Delta X^k)$ $-Y^k$ is reached. Among the three steps, the optimum of (9) is found by alternatively solving Y, Λ and ΔX respectively:

$$Y^{k+1}=\arg\min_Y\left\|Y\right\|_*+\frac{2\rho+\sigma_k}{2}\left\|Y-\varsigma\right\|_F^2 = D_{1/(2\rho+\sigma_k)}(\varsigma)$$

$$\Lambda_{k+1}=\arg\min_\Lambda \rho\left\|Y_k-A_0-A\Lambda\right\|_F^2 = A^T\left(Y_k-A_0\right)$$

$$\Delta X_{k+1}=\arg\min_{\Delta X}\frac{\sigma_k}{2}\left\|g(\Delta X)-Y_k+Z_k/\sigma_k\right\|_F^2 = \sum_i J_i^T\left(Y_k-\sum_i T(I_i;x_i)e_i^T-Z_k/\sigma_k\right)e_i e_i^T$$

$$\varsigma=\frac{1}{2\rho+\sigma_k}\left[2\rho\left(A_0+A\Lambda_k\right)+\sigma_k g\left(\Delta X_k\right)+Z_k\right]$$

where D is the singular value thresholding operator defined in [25], and ΔX is the matrix form of the increments Δx_i.

When the inner loop (the three update steps) converges, we update the x_i in (7) and obtain the final result of (6) until its convergence is reached.

4.5. Re-representation of *G* and *W*

In this paper, we consider the global transformation G as affine (translation, in-plane rotation and scaling), rather than perspective because the out-of-plane rotation of a head has been addressed by the PCs in the statistic shape model.

Moreover, we parameterize G in a form which operates on a shape rather than on the coordinates of its landmarks. Specifically, the shape after global transformation is:

$$G(s;q)=s_0+\sum_{i=1}^4 q_i s_i^*$$

where s is the shape to be transformed; q_i are the parameters of the affine transformation G; and s_i^* are the extended basis vectors defined as:

$$s_1^*=\left(x_1,y_1,...,x_n,y_n\right)^T,\quad s_2^*=\left(-y_1,x_1,...,-y_n,x_n\right)^T$$
$$s_3^*=\left(1,0,...,1,0\right)^T,\qquad s_4^*=\left(0,1,...,0,1\right)^T$$

It has been demonstrated in [14] that this representation is equivalent to that operating on the 2D coordinates. However, its advantage is obvious: it is consistent to the shape model, so that the global transformation is parameterized together with the shape deformation to form an n+4

dimensional vector. We therefore denote the concatenation of p and q by x, and denote the composition of G and W by T.

4.6. Discussion

It is worth noticing that, while ρ approaches infinity, the objective function (6) reduces to the problem of "Independent AAM"; on the other hand, when ρ approaches 0, it can be perceived as an extension of [7] from rigid alignment to non-rigid. The extension is in a similar spirit as [6], yet the difference is that instead of using B-spline as the authors in [6] did, we employ a learned deformable shape model for non-rigid deformation.

As setting a proper ρ is important for the success of our method, to determine an optimal value is not straightforward. As mentioned in Section 4.1, the value of ρ is dependent on the accuracy of the generic appearance model and the correlation of the images. Here we provide a coarse estimation of its value

$$\rho = C \cdot \left\| Y^0 \right\|_* / \left\| Y^0 - A_0 - A\Lambda^0 \right\|_F$$

in a sense that it grows with the uncorrelation of the images and descends with the inaccuracy of the generic appearance model. Here Y^0 and Λ^0 are the initial state of Y and Λ, and C is an empirically determined constant which is set to be 1 in the following experiments.

Moreover, although the formulation (5) of our problem appears similar to that in [7], it is worth mentioning the differences: First, we address the non-rigid alignment problem, which is more challenging and finds applications in a variety of more ambitious scenarios, for example those mentioned in Section 2. Second, directly extending the application of rank minimization from rigid to non-rigid circumstance by merely increasing the dimension of parameters does not give plausible results. The reason is that the aligned shapes do not converge to human facial contours since the generic human face assumption is ignored. Third, while the authors directly minimize the rank of the aligned faces, we employ it as a regularization term. Last but not least, we employ the ALM method which enjoys the advantages listed in Section 4.4.

5. Experiments

In this section, we specify our experiments and compare the obtained results with other methods.

5.1. Settings

In our experiment, we train a generic AAM model from the public available IMM database [27] which consists of a total of 240 labeled images of 40 persons. The images are taken under controlled laboratory conditions with varying head poses and facial expressions. During the learning stage, we perform PCA on the training shapes and appearances, and retain 95% of their variations, resulting in 21 PCs and 98 PCs for each of the two models.

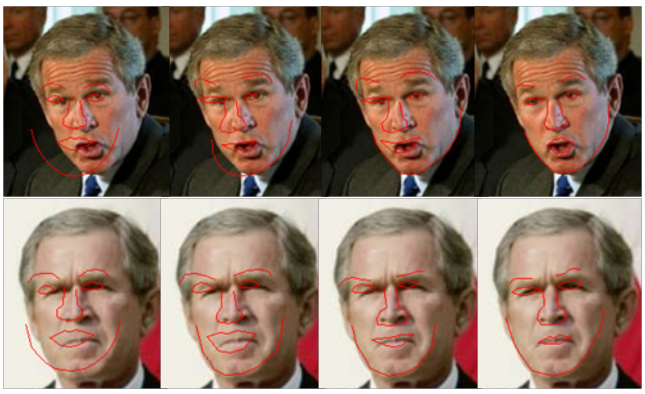


Figure 5. Iterative shape update in joint alignment of "Bush"

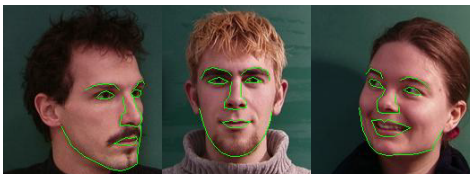


Figure 6. Examples of the training set [27]

We test the algorithms on the LFW database [28], which consists of photos of 20 persons retrieved from the internet. It is worth mentioning that, the images are taken under wild conditions with challenging and (some) extreme poses, expressions and illuminations. Moreover, the faces in the images are drastically different from the training ones. These two facts pose a steep challenge for existing methods: the appearance model learned from IMM [27] is substantially inaccurate in modeling the faces in LFW [28], as it has been shown in Fig.4. However, it is important for a good non-rigid alignment algorithm to be applicable in such a practical scenario.

We use publicly available software [29, 30] for the implementation of [13] and [14], with all parameters set by default. The only modifications we make are that, for a fair comparison purpose, we modify the shape initialization functions to make sure they initialize the same shapes as ours does: the shapes are initialized to be the reference shape s_0, with their positions and sizes adjusted according to the face detection results.

5.2. Results and Discussions

In Fig.7 we show a part of our results (4th row) in comparison with: Independent AAM (top), RASL (2nd row) and their composition (3rd row). It is worth mentioning that in the visualization of RASL results we try our best at tuning the size of the reference shape and its offset from the

Top row: Independent AAM [13]; *second row*: RASL [7]; *third row*: RASL + AAM; *fourth row*: our method; *bottom row*: ground-truths

Figure 7. Alignment results on "Sharon" in comparison with other work

face rectangles returned by the RASL algorithm. Therefore, we can perceive the observed errors as that a rigid alignment technique generates in a non-rigid task. We manually labeled 5 arbitrarily selected image sets consisting of totally 157 images, and show some of them in the bottom row. For more results, please refer to our supplementary material.

Table 1: Mean point-point error (pixels)

	AAM	RASL+AAM	Ours	Ground
Obama	10.46	9.54	**1.64**	0.91
Bush	10.14	10.25	**3.52**	1.89
Sharon	9.92	8.39	**2.45**	2.11
Rumsfeld	9.61	9.10	**4.00**	1.79
Ashcroft	10.22	5.66	**3.24**	1.84
Average	10.07	8.59	**2.97**	1.71

To quantify the error, we record in Table 1 the mean point-point error in terms of Root Mean Square (RMS). Notice that the 4th column refers to the error of fitting a set of learned "ground-truth" person-specific AAM models to each of the input image sets respectively. It can be regarded as the minimum error that an algorithm can achieve on the dataset in an idealistic scenario. In addition, to evaluate the robustness against the initialization error of the algorithms, we systematically perturb the ground-truths by: ± 10-pixels in x/y translation, $\pm \pi/8$ in rotation and ± 0.2 in scaling, and count convergence of a fit if its error is less than 4 pixels. The average convergence rate of 10 trials is given in Tab. 2.

Table 2: Convergence rate (%)

AAM	AAM(IC)	RASL+AAM	Ours
43.81	50.29	44.66	**90.86**

As can be observed, our method performs consistently better than the others in the joint non-rigid alignment task. It is obvious and worth noticing that, conventional AAMs still get stuck into unwanted solutions although an accurate pose initialization has been given by RASL. The main reason is due to, as discussed in Section 3.3, the inaccuracy of the appearance model and biased gradient directions.

Besides the fact that the overwhelming majority of our results are good, there are also negative examples under challenging conditions: those with exaggerated expressions, or with focusing blur and large outliers. For these images, the assumptions made in Section 4.1 are violated due to occlusions and shadows. In the right hand of Fig.9, we show some of these examples. Notice that, as long as they are substantially minor in number, our method can still reliably identify the appearance of the input faces and generate good results on the normal ones, such as those

Figure 8. More results of the proposed method

shown in Fig.7 and Fig.8.

6. Conclusion and Future Work

In this paper, we propose a model-based method for joint alignment of multiple images of an object undergoing a variety of appearance variations. The method does not rely on an accurate appearance model which is learned from training data as conventional ones do. It simultaneously identifies the appearance of the interested object while bringing the images into alignment. Experiments on wild conditioned dataset demonstrate the effectiveness of the proposed method.

On the other hand, our ongoing work involves performance improvement in more challenging circumstances where outliers are present and images are degraded.

Figure 9. Some results on the extreme cases

References

[1] E. Learned-Miller. Data driven image models through continuous joint alignment. *PAMI*, 2006.

[2] A. Vedaldi, G. Guidi, and S.Soatto, Joint data alignment up to (lossy) transformations, *CVPR*, 2008

[3] Mark. Cox, et al. Least Squares Congealing for Unsupervised Alignment of Images, *CVPR*, 2008

[4] Mark. Cox, et al. Least-Squares Congealing for Large Numbers of Images, *ICCV*, 2009

[5] G.B. Huang, V. Jain, E. Learned-Miller, Unsupervised Joint Alignment of Complex Images, *ICCV*, 2007

[6] Hongjun Jia, et al. ABSORB: Atlas Building by Self-Organized Registration and Bundling, *CVPR*, 2010

[7] Y. Peng, et al. RASL: Robust Alignment by Sparse and Low-rank Decomposition for Linearly Correlated Images, *CVPR*, 2010

[8] Z.Cao, Q.Yin, J.Sun, X.Tang, Face Recognition with Learning-based Descriptor, *CVPR*, 2010

[9] D. Bitouk, et al. Face Swapping: Automatically Replacing Faces in Photographs, *SIGGRAPH*, 2008

[10] J. Zhu, et al. Automatic 3D Face Modeling Using 2D Active Appearance Models, *Pacific Graphics*, 2005

[11] S.K. Balci, Free-form B-Spline Deformation Model For Groupwise Registration. *MICCAI*, 2007

[12] V. Blanz T. Vetter, A Morphable Model For The Synthesis Of 3D Faces, *SIGGRAPH*, 1999

[13] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active Appearance Models, *ECCV*, 1998

[14] I. Matthews and S. Baker, Active Appearance Models Revisited, *IJCV*, 2004

[15] R. Gross,et al. Generic vs. Person Specific Active Appearance Models, *J. of Image and Vision Computing*, 2005

[16] S.J. Lee, et al. A Comparative Study of Facial Appearance Modeling Methods for Active Appearance Models, *PR*, 2010

[17] Z. Xu, H. Chen, S.C. Zhu, A Hierarchical Compositional Model for Face Representation and Sketching, *PAMI* 2008

[18] M.H. Nguyen, F. Torre, Metric Learning for Image Alignment, *IJCV*, 2009

[19] T.F. Cootes and C.J. Taylor, An Algorithm for Tuning an Active Appearance Model to New Data, *BMVC*, 2006

[20] G. Papandreou, et al. Adaptive and Constrained Algorithms for Inverse Compositional AAM Fitting, *CVPR*, 2008

[21] R. Basri, et al. Lambertian Reflectance and Linear Subspaces, *IEEE Trans. on PAMI*, 2003

[22] E.J.Candes, et al. Exact Matrix Completion via Convex Optimization, *Found. of Computational Mathematics*, 2008

[23] E.J.Candes and Y. Plan, Matrix Completion with Noise, *IEEE Proceedings* 2009

[24] J. Wright, Y. Peng, and Y. Ma, Robust Principal Component Analysis, *NIPS* 2009

[25] J. Cai, E.J.Candes, and Z. Wen, A Singular Value Thresholding Algorithm for Matrix Completion, *SIAM J. on Optimization* 2008

[26] Z. Lin, M. Chen, L. Wu and Y. Ma, The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices, *UIUC Report* 2009

[27] M.M. Nordstrom, The IMM Face Database - An Annotated Dataset of 240 Face Images, *Technical Report*, 2004

[28] G.B. Huang, et al. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, *Technical Report*, 2007

[29] M.B. Stegmann, et al. A Flexible Appearance Modeling Environment, *IEEE Trans. on Medical Imaging*, 2003

[30] http://code.google.com/p/aam-library/