Rank Minimization across Appearance and Shape for AAM Ensemble Fitting

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

[1]Queensland University of Technology
[1]{x2.cheng,s.sridharan}@qut.edu.au
[2]The Commonwealth Scientific and Industrial Research Organization (CSIRO)
[2]{jason.saragih,simon.lucey}@csiro.au

Abstract

Active Appearance Models (AAMs) employ a paradigm of inverting a synthesis model of how an object can vary in terms of shape and appearance. As a result, the ability of AAMs to register an unseen object image is intrinsically linked to two factors. First, how well the synthesis model can reconstruct the object image. Second, the degrees of freedom in the model. Fewer degrees of freedom yield a higher likelihood of good fitting performance. In this paper we look at how these seemingly contrasting factors can complement one another for the problem of AAM fitting of an ensemble of images stemming from a constrained set (e.g. an ensemble of face images of the same person).

1. Introduction

Active Appearance Models (AAMs) employ linear models of shape and appearance. However, AAMs are essentially non-linear parametric models of pixel intensities. Fitting an AAM to an image is therefore inherently a non-linear optimisation problem. A well known issue in non-linear optimisation, and thus AAMs, is local minima. An obvious strategy for dealing with local minima is to reduce the degrees of freedom.

Unfortunately, many of the objects that AAMs are traditionally aimed at (such as human faces, organs in medical imaging, etc.) have considerable variation in both shape and appearance. Gross et al. [6] demonstrated this problem explicitly for the task of non-rigid face fitting. Specifically, Gross et al. showed that: (i) person specific AAMs substantially outperform a generic AAM (i.e. models trained across many subjects), and (ii) this disparity in performance stems from the high degree of freedom of the generic AAM.

The Problem: AAM fitting is typically applied to an ensemble of images stemming from a similar source. A prime example of a similar source is in a temporal image sequence stemming from the same object (e.g. a single face). All the images in the sequence are different, but the overall variation in the specific object is quite small (e.g. expressions, pose, illumination variation for a single face). As a result the variation for the object can be modelled quite compactly in terms of a linear shape and appearance basis. In fact, the sequence does not need to be ordered in any particular causal manner, so we instead use the term *ensemble*.

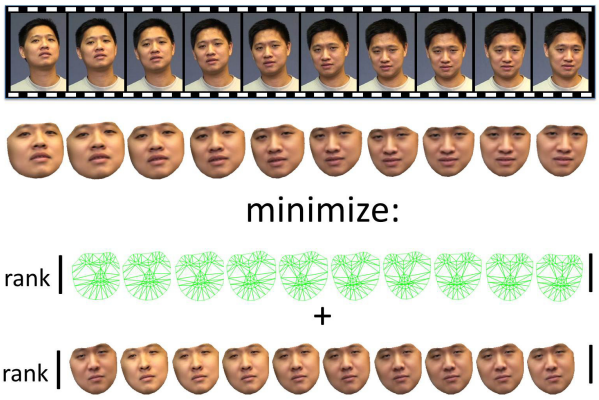


Figure 1. The proposed method simultaneously fits generic AAMs to images in an ensemble by constraining the shape and appearance variations with rank minimization. As a result the ensemble-specific AAM is determined with the ensemble's specific shape and appearance variations.

It is obvious that if one has a priori knowledge of this ensemble's specific shape and appearance basis, one could apply standard AAM fitting methods. Unfortunately, one rarely has such knowledge as an external agent would need to manually register the images in the ensemble to construct the AAM, thus defeating the purpose of the entire AAM fitting exercise. Instead one often resorts to generic AAM methods which result in sub-optimal performance. We de-

fine a *generic* AAM, as a model whose shape and appearance basis has been estimated to model all instances of the object being modelled (e.g. faces of all the population). We define an *ensemble-specific* AAM, as a model whose shape and appearance basis has been estimated to compactly model a specific instance of the object being modeled (e.g. single face in the ensemble).

Contributions: In this paper we explore the ambitious problem of automatically determining an ensemble-specific AAM directly from the ensemble in an unsupervised manner. We draw inspiration from recent works in unsupervised image ensemble alignment [4,5,11,14,17], specifically Robust Alignment by Spare and Low-rank (RASL) decomposition [14]. RASL attempts to align images in an ensemble by assuming that the aligned image ensemble is compact in terms of image variation. However, RASL is not able to manage either deformable objects, nor prior about an object (e.g. generic AAM) in its current framework. In this paper, we propose a RASL inspired generic AAM fitting algorithm for image ensembles. Specifically, we make three contributions in this paper:

- We show how the ensemble-specific AAM can be determined by applying a rank minimization strategy to shape and appearance variations in conjunction with the standard AAM fitting objective function (Section 4).
- We empirically show that in the specific application of face fitting, applying rank constraints on shape and appearance variations together yield notable better performance than constraining appearance variation alone (Section 6).
- We show that the ensemble-specific AAM determined by the proposed method has lower degrees of freedom than the generic AAM. Further, the ensemble-specific AAM is capable of being applied to additional images of the same instance through canonical efficient AAM fitting methods (Section 6).

Related Work: There are many methods proposed for non-rigid image ensemble alignment [2, 16, 19]. Most notably, Zhao et al. proposed a RASL inspired generic AAM fitting approach [19]. Their approach simultaneously fits generic AAMs to all images in the ensemble by constraining the compactness of the aligned appearances. However, their method has some limitations: (i) the degrees of freedom with respect to shape is not considered (they only considered appearance); (ii) their approach is not robust to partial occlusions as they employ an outlier sensitive error function (L2 norm square); and (iii) their approach has an inherent limitation when applied to large scale problems since all images in the ensemble have to attend the alignment simultaneously.

Mathematical Notations: Vectors are always represented in lower-case bold (e.g., $\mathbf{a}$). Matrices are always expressed in upper-case bold (e.g., $\mathbf{A}$). Scalars in lower-case (e.g. a). Images in this paper shall always be expressed in capitalized form A. Warp functions $\mathcal{W}(\mathbf{x};\mathbf{p})$ will be used throughout this paper to denote a warping of a $2D$ coordinate vector $\mathbf{x} = [x, y]^T$ by a warp parameter vector $\mathbf{p} \in \mathbb{R}^P$, where P is the number of warp parameters, back to a fixed base coordinate system. This base coordinate system is defined when $\mathbf{p} = \mathbf{0}$ such that $\mathcal{W}(\mathbf{x};\mathbf{p}) = \mathbf{x}$. An abuse of notation is entertained in this paper for when an image I is warped by the warp parameter vector $\mathbf{p}$, such that $I(\mathbf{p}) = [I(\mathcal{W}(\mathbf{x}_1;\mathbf{p})), \ldots, I(\mathcal{W}(\mathbf{x}_D;\mathbf{p}))]^T$. In this instance $I(\mathbf{p})$ is a D dimensional vector of image intensities, where D denotes the number of discrete coordinates in the base coordinate system. The Jacobian matrix $\mathbf{J} = \frac{\partial I(\mathbf{p})}{\partial \mathbf{p}}$ of an image $I(\mathbf{p})$ is used frequently through out this paper. This $D \times P$ matrix is formed by combining image gradients of $I(\mathbf{p})$ with the Jacobian of the warp function $\mathcal{W}(\mathbf{x};\mathbf{p})$, more details on the formation of this matrix can be found in [13].

2. AAMs

Active appearance models (AAMs) [3, 13] are usually constructed from a set of training images with the AAM mesh vertices hand-labelled on them. The training mesh vertices are first aligned with Procrustes Analysis. Then principal component analysis (PCA) is used to build a 2D linear model of shape variation. The 2D shape $\mathbf{s} = (x_1, y_1, \ldots, x_V, y_V)^T$ can be represented as a base shape $\mathbf{s}_0$ plus a linear combination of P shape vectors $\mathbf{s}_i$,

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{P} p_i \mathbf{s}_i = \mathbf{s}_0 + \mathbf{\Phi}\mathbf{p}, \tag{1}$$

where $\mathbf{p} = [p_1, \ldots, p_P]^T$ is the shape parameter vector and $\mathbf{\Phi} = [\mathbf{s}_1, \ldots, \mathbf{s}_P]^T$ is the matrix of concatenated shape vectors. The AAM model of appearance variation is obtained by first warping all the training images onto the mean shape and then applying PCA on the shape normalized appearance images. The appearance of an AAM $A(\mathbf{0})$ is an image vector defined over the pixels $\mathbf{x} \in \mathbf{s}_0$ when $\mathbf{p} = \mathbf{0}$. The appearance $A_{\boldsymbol{\lambda}}(\mathbf{0})$ can be represented as a mean appearance $A_0(\mathbf{0})$ plus a linear combination of K orthonormal appearance vectors $A_j(\mathbf{0})$,

$$A_{\boldsymbol{\lambda}}(\mathbf{0}) = A_0(\mathbf{0}) + \sum_{j=1}^{K} \lambda_j A_j(\mathbf{0}) = A_0(\mathbf{0}) + \mathbf{A}\boldsymbol{\lambda}, \tag{2}$$

where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]^T$ is the appearance parameter vector and $\mathbf{A} = [A_1(\mathbf{0}), \ldots, A_K(\mathbf{0})]$ is the matrix of concatenated appearance vectors.

To fit the predefined AAMs to the image, one can use the fitting algorithm based on the Lucas & Kanade (LK) algorithm [12]. In this approach one can pose AAM fitting as minimizing the following objective function,

$$\arg\min_{\mathbf{p},\boldsymbol{\lambda}} \parallel I(\mathbf{p}) - A_0(\mathbf{0}) - \mathbf{A}\boldsymbol{\lambda} \parallel_2^2 \tag{3}$$

where $I(\mathbf{p})$ represents the warped input image using the warp specified by the parameters $\mathbf{p}$. The central task of this objective function is to find the shape $\mathbf{p}$ and appearance $\boldsymbol{\lambda}$ that minimizes the sum of squared distances (SSD) between the warped input image and the AAM. Since the relationship between the warp parameters $\mathbf{p}$ and the warped image $I(\mathbf{p})$ is non-linear, a first order Taylor series linear approximation, $I(\mathbf{p} + \Delta\mathbf{p}) \approx I(\mathbf{p}) + \mathbf{J}\Delta\mathbf{p}$, is employed, where $\mathbf{J}$ stands for the image Jacobian matrix.

3. RASL

Robust Alignment by Sparse and Low-rank (RASL) decomposition [14] method was built based on an assumption that the warped image ensemble matrix $\mathbf{D}(\mathbf{P}) = [(I_1(\mathbf{p}_1), \ldots, I_F(\mathbf{p}_F)]$ is of low rank and the image errors are sparsely distributed [14], where $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_F]$ is the matrix of warp parameters for all F frames in the image ensemble. RASL is an extension of the earlier work of Robust Principal Component Analysis [18]. The central idea is to find the transformation between the original image and the warped image ensemble matrix by minimizing the rank of matrix $\mathbf{L}$ and the number of non-zero errors $\mathbf{E}$,

$$\begin{aligned} \arg\min_{\mathbf{L},\mathbf{E},\mathbf{P}} \quad & \mathrm{rank}(\mathbf{L}) + \lambda||\mathbf{E}||_0 \\ \text{s.t.} \quad & \mathbf{D}(\mathbf{P}) = \mathbf{L} + \mathbf{E}, \end{aligned} \tag{4}$$

where $\mathbf{L}$ and $\mathbf{E}$ are matrices with same dimension of $\mathbf{D}$. The authors in [14] relaxed the objective convexity by replacing $\mathrm{rank}(\cdot)$ and $||\cdot||_0$ with their convex approximations, namely the nuclear norm $||\cdot||_*$ and $\mathcal{L}1$-norm $||\cdot||_1$ respectively. This results in the following objective,

$$\begin{aligned} \arg\min_{\mathbf{L},\mathbf{E},\boldsymbol{\Delta}\mathbf{P}} \quad & ||\mathbf{L}||_* + \lambda||\mathbf{E}||_1 \\ \text{s.t.} \quad & \mathbf{D}(\mathbf{P}) + \sum_{i=1}^{F} \mathbf{J}_i \Delta\mathbf{P} \epsilon_i \epsilon_i^{\mathbf{T}} = \mathbf{L} + \mathbf{E} \ . \end{aligned} \tag{5}$$

A first order Taylor series linear approximation, $\mathbf{D}(\mathbf{P} + \Delta\mathbf{P}) \approx \mathbf{D}(\mathbf{P}) + \sum_{i=1}^{F} \mathbf{J}_i \Delta\mathbf{P} \epsilon_i \epsilon_i^{\mathbf{T}}$, is employed in this equation, where $\mathbf{J}_i$ is the image Jacobian matrix of the i^{th} image, ϵ_i is the $F \times 1$ standard basis vector (all elements in this vector are zeros except i^{th} element is one), $\Delta\mathbf{P}$ is the increment update of the warp parameter $\mathbf{P}$.

4. Joint Face Ensemble Alignment

Earlier Work by Zhao et al.: In contrast to the conventional pair-wise image alignment methods such as Lucas-Kanade inspired AAMs [3, 13], we proposed to fit an AAM to all images in an image ensemble simultaneously. The most recent and related work was proposed by Zhao et al. [19]. Their approach employs an AAMs objective term to regularize the nuclear norm optimization, to ensure the aligned facial appearances are within the variations defined by a generic AAM,

$$\arg\min_{\mathbf{P},\boldsymbol{\Lambda}} \quad \mathrm{rank}(\mathbf{D}(\mathbf{P})) + \lambda||\mathbf{D}(\mathbf{P}) - \mathbf{A}_0 - \mathbf{A}\boldsymbol{\Lambda}||_2^2, \tag{6}$$

where $\mathbf{D}(\mathbf{P})$ is the facial appearances transformed into the reference shape frame, $\mathbf{A}_0$ is the matrix composed of replicas of the reference appearance $A_0(\mathbf{0})$, and each column of the matrix $\boldsymbol{\Lambda}$ is a vector of appearance coefficients of that particular frame, $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_F]$. In [19], the generic appearance model $\mathbf{A}$ employed in their implementation and experiment was formed by 98 appearance eigenvectors.

Our Objective: In [6], Gross et al. showed that in the task of generic AAMs fitting, the shape component is the main cause of the reduced fitting robustness. In this paper, we propose an improved method which offers better AAM fitting accuracy and robustness by, (i) applying rank constraints on shape and appearance variations at the same time; (ii) exploiting robust L0 norm function instead of the L2 norm square to improve the robustness to image outliers. We formulate the problem as,

$$\begin{aligned} \arg\min_{\mathbf{P},\boldsymbol{\Lambda}} \quad & \mathrm{rank}(\mathbf{A}\boldsymbol{\Lambda}) + \lambda_1 \mathrm{rank}(\boldsymbol{\Phi}\mathbf{P}) \\ & + \lambda_2||\mathbf{D}(\mathbf{P}) - \mathbf{A}_0 - \mathbf{A}\boldsymbol{\Lambda}||_0, \end{aligned} \tag{7}$$

where $\mathbf{A}\boldsymbol{\Lambda}$ represents the appearance variations of all images in the ensemble, and $\boldsymbol{\Phi}\mathbf{P}$ represents the shape variations, λ_1 and λ_2 are weights. Note that in contrast to Equation 6, we apply low rank constraint on $\mathbf{A}\boldsymbol{\Lambda}$ instead of $\mathbf{D}(\mathbf{P})$. This is because $\mathbf{D}(\mathbf{P})$ is no longer low rank because of the existence of the image outliers (e.g. occlusion, shadow). The proposed method searches for the parameters $\mathbf{P}$, $\boldsymbol{\Lambda}$ such that the appearance and shape variations are most compact. $||\cdot||_0$ is the number of non-zeros elements, this special norm term is preferred instead of the conventional $||\cdot||_2^2$ since it is robust to image outliers [14].

The convexity of our objective function is relaxed using the same methodology as described in [14, 18]. This results in the following objective,

$$\arg\min_{\mathbf{P},\mathbf{\Lambda}} \quad ||\mathbf{A\Lambda}||_* + \lambda_1||\mathbf{\Phi P}||_* \quad (8)$$
$$+\lambda_2||\mathbf{D}(\mathbf{P}) - \mathbf{A}_0 - \mathbf{A\Lambda}||_1.$$

5. ADMM Optimization

Reformulation: It has been proven in [14] that the Alternating Direction Method of Multipliers (ADMM) [1] is extremely efficient to solve objective function which includes L1 norm $||\cdot||_1$ or nuclear norm $||\cdot||_*$. To solve our convex objective function using ADMM, we reformulate the objective of Equation 8 to,

$$\begin{aligned} \arg\min_{\mathbf{\Delta P},\mathbf{\Lambda}} \quad & ||\mathbf{G}||_* + \lambda_1||\mathbf{X}||_* + \lambda_2||\mathbf{E}||_1 \quad (9) \\ \text{s.t.} \quad & \mathbf{G} = \mathbf{\Lambda}, \\ & \mathbf{X} = \mathbf{\Phi P} + \mathbf{\Phi}\Delta\mathbf{P}, \\ & \mathbf{E} = \mathbf{D}(\mathbf{P}) + \sum_{i=1}^{F} \mathbf{J}_i \Delta\mathbf{P}\epsilon_i\epsilon_i^{\mathbf{T}} - \mathbf{A}_0 - \mathbf{A\Lambda}, \end{aligned}$$

where $\mathbf{G}$, $\mathbf{X}$ and $\mathbf{E}$ are auxiliary variables to allow us to solve the objective using ADMM and the efficient soft-threshold methods, $\mathbf{G}$ represents the appearance coefficients (same as $\mathbf{\Lambda}$). $\mathbf{E}$ represents the errors between the current alignment and the estimated facial appearance. $\mathbf{X}$ represents the shape with the updated shape coefficients $\mathbf{P}+\Delta\mathbf{P}$. Note in this formulation, we applied nuclear norm to the appearance coefficients directly instead of the appearance variations $\mathbf{A\Lambda}$. This is because the linear appearance model $\mathbf{A}$ estimated by Principal Component Analysis is orthogonal, then we have $||\mathbf{A\Lambda}||_* = ||\mathbf{\Lambda}||_*$.

ADMM Optimization: To solve the objective function of Equation 9, we rewrote the objective function in Augmented Lagrangian form, in which the equality constraints are appended into the objective function. The Augmented Lagrangian function can then be optimized by solving each of the variables alternately until converges. We write our Augmented Lagrangian Function in the scaled form [1] as,

$$\begin{aligned} &\mathcal{L}(\mathbf{G},\mathbf{E},\mathbf{X},\mathbf{\Lambda},\mathbf{\Delta P},\boldsymbol{\xi_1},\boldsymbol{\xi_2},\boldsymbol{\xi_3}) = \\ &||\mathbf{G}||_* + \lambda_1||\mathbf{X}||_* + \lambda_2||\mathbf{E}||_1 + \frac{\mu_1}{2}||\mathbf{G} - \mathbf{\Lambda} + \frac{1}{\mu_1}\boldsymbol{\xi}_1||_2^2 \\ &+ \frac{\mu_2}{2}||\mathbf{X} - \mathbf{\Phi P} - \mathbf{\Phi}\Delta\mathbf{P} + \frac{1}{\mu_2}\boldsymbol{\xi}_2||_2^2 + \frac{\mu_3}{2}||\mathbf{\Gamma}||_2^2, \quad (10) \end{aligned}$$

where $\mathbf{\Gamma} = \mathbf{D}(\mathbf{P}) + \sum_{i=1}^{F} \mathbf{J}_i\Delta\mathbf{P}\epsilon_i\epsilon_i^{\mathbf{T}} - \mathbf{A}_0 - \mathbf{A\Lambda} - \mathbf{E} + \frac{1}{\mu_3}\boldsymbol{\xi}_3$, $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_3$ are the Lagrangian multipliers, μ_1, μ_2 and μ_3 are positive scalers. Each variable of $\mathbf{\Delta P}, \mathbf{\Lambda}, \mathbf{E}, \mathbf{X}$ and $\mathbf{G}$ can be determined through a Gauss-Seidel style alternation strategy as described in Algorithm 1.

Algorithm 1 Alterative optimization of ADMM

while NOT CONVERGED **do**
  Update $\mathbf{G}$: $\arg\min_{\mathbf{G}} \mathcal{L}(\cdots)$,
  Update $\mathbf{X}$: $\arg\min_{\mathbf{X}} \mathcal{L}(\cdots)$,
  Update $\mathbf{E}$: $\arg\min_{\mathbf{E}} \mathcal{L}(\cdots)$,
  Update $\mathbf{\Lambda}, \mathbf{\Delta P}$: $\arg\min_{\mathbf{\Lambda},\mathbf{\Delta P}} \mathcal{L}(\cdots)$,
  Update $\boldsymbol{\xi}_1$: $\boldsymbol{\xi}_1 + \mu_1(\mathbf{G} - \mathbf{\Lambda})$,
  Update $\boldsymbol{\xi}_2$: $\boldsymbol{\xi}_2 + \mu_2(\mathbf{X} - \mathbf{\Phi P} - \mathbf{\Phi\Delta P})$,
  Update $\boldsymbol{\xi}_3$: $\boldsymbol{\xi}_3 + \mu_3(\mathbf{\Gamma} - \frac{1}{\mu_3}\boldsymbol{\xi}_3)$,
  Update μ_i: $a \cdot \mu_i$.
end while

Here a is an incremental factor for the scalars μ_1, μ_2 and μ_3. The value of a that yields the best efficiency was experimentally found to be $a = 1.25$. The initial values of $\boldsymbol{\xi}_1^0$, $\boldsymbol{\xi}_2^0$, $\boldsymbol{\xi}_3^0$, μ_1^0, μ_2^0, μ_3^0 were selected using the same methodology as described in [18].

Efficient Sub-Problems: ADMM is extremely efficient as it enables one to break a complex objective into a sequence of efficient sub-problems. The updates of $\mathbf{G}, \mathbf{E}$ and $\mathbf{X}$ can be solved efficiently by the soft-threshold methods as described in [1, 14], the appearance coefficients $\mathbf{\Lambda}$ and the incremental shape coefficients $\mathbf{\Delta P}$ are updated as,

$$\begin{aligned} [\mathbf{\Lambda},\mathbf{\Delta P}] = \arg\min_{\mathbf{\Lambda},\mathbf{\Delta P}} & \frac{\mu_1}{2}||\mathbf{G} - \mathbf{\Lambda} + \frac{1}{\mu_1}\boldsymbol{\xi}_1||_2^2 \\ + \frac{\mu_2}{2}||\mathbf{X} - \mathbf{\Phi P} - \mathbf{\Phi\Delta P} + \frac{1}{\mu_2}\boldsymbol{\xi}_2||_2^2 &+ \frac{\mu_3}{2}||\mathbf{\Gamma}||_2^2 \ . \quad (11) \end{aligned}$$

The forward compositional "project-out" [13] algorithm is used to solve Equation 11. This algorithm was used because it is extremely efficient, especially when it is used as an iterative update in the loop of ADMM. In this method, $\mathbf{\Gamma}$ is decomposed into two terms,

$$||\mathbf{\Gamma}||_2^2 = ||\mathbf{\Gamma}||_{\text{span}(\mathbf{A})\perp}^2 + ||\mathbf{\Gamma}||_{\text{span}(\mathbf{A})}^2, \quad (12)$$

where $||\cdot||_{\text{O}}^2$ denotes the square of L2 norm of the vector projected into the linear subspace of O, $\text{span}(\mathbf{A})$ is the subspace spanned by the appearance basis $\mathbf{A}$, and $\text{span}(\mathbf{A}) \perp$ is its its orthogonal complement. Note in the first term, the norm function only considers the components of vectors in the orthogonal complement of $\text{span}(\mathbf{A})$. This term is thus invariant to $\mathbf{\Lambda}$. Then we have,

$$\begin{aligned} \mathbf{\Delta P} = \arg\min_{\mathbf{\Delta P}} & \frac{\mu_2}{2}||\mathbf{X} - \mathbf{\Phi P} - \mathbf{\Phi\Delta P} + \frac{1}{\mu_2}\boldsymbol{\xi}_2||_2^2 \\ &+ \frac{\mu_3}{2}||\mathbf{\Gamma}||_{\text{span}(\mathbf{A})\perp}^2. \quad (13) \end{aligned}$$

The Jacobians $\mathbf{J}_i$ can be determined as described in [13]. The appearance coefficients $\mathbf{\Lambda}$ can then be determined as a

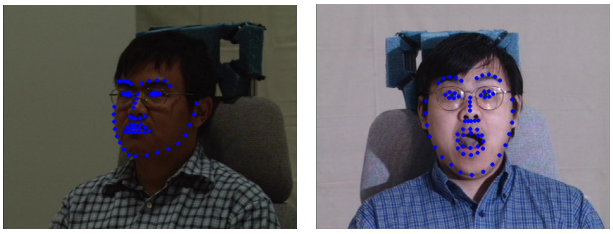


Figure 2. MultiPIE training samples with varying head poses, facial expressions and illumination conditions with the facial landmark annotation.

Least Square Problem,

$$\mathbf{\Lambda} = \arg\min_{\mathbf{\Lambda}} \frac{\mu_1}{2}||\mathbf{G} - \mathbf{\Lambda} + \frac{1}{\mu_1}\boldsymbol{\xi}_1||_2^2 + \frac{\mu_3}{2}||\mathbf{\Gamma}||_2^2 \quad . \quad (14)$$

6. Experiments

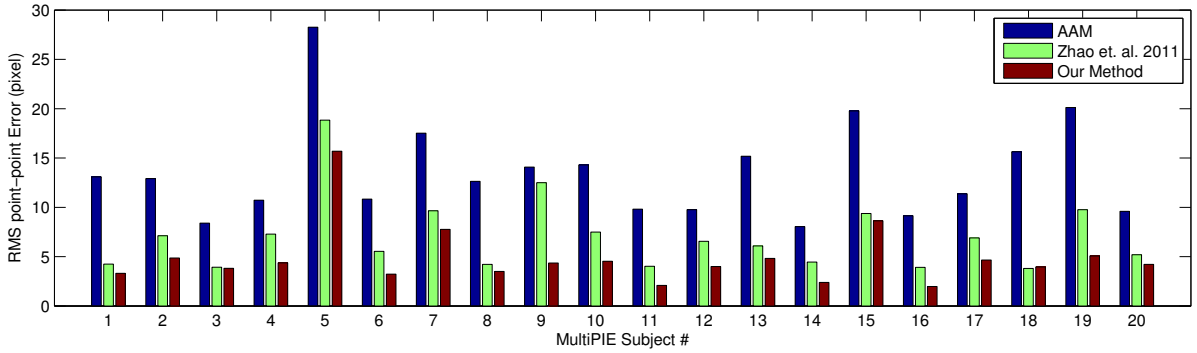
This section describes our experiments on several publicly available image databases and video databases with varying image conditions.

Implementation/Setup: The generic AAM applied in the experiments was trained using the MultiPIE database [7] and Cohn-Kanade database [9]. The MultiPIE training samples (as demonstrated in Figure 2) include identities from subject 21 to subject 346 with different head poses, facial expressions and illumination variations (subject 1 to subject 20 were reserved for testing). The obtained AAM includes 295 appearance basis vectors and 20 shape basis vectors (98% of the variations). In the implementation of the proposed method, the weight, λ_1 was selected using the same strategy as proposed in [14], $\lambda_1 = 1/\sqrt{D}$, where D is the number of pixels in each appearance basis (30,000 in our model). λ_2 was selected by $\lambda_2 = \sqrt{\frac{D}{V}}$, where V is the number of landmark points (66 in our model). In our implementation of [19], we selected the default weight $\lambda = ||\mathbf{D}(\mathbf{P}^0)||_*/||\mathbf{D}(\mathbf{P}^0) - \mathbf{A}_0 - \mathbf{\Psi}\mathbf{\Lambda}^0||_2^2$ as proposed by the authors. All the RMS registration errors in our experiments were determined in the reference shape system defined by AAM, in our AAM the size of face image is 116×113 pixels. The Constrained Local Models (CLMs) evaluated in our experiments was implemented and published by the authors of [15]. The CLMs were learnt from MultiPIE database.

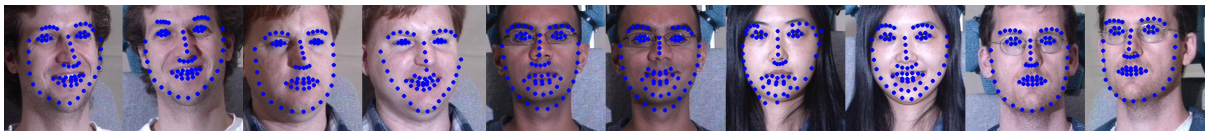
MultiPIE Database: This section describes our experiments on the MultiPIE [7] database. The first 20 MultiPIE subjects were sampled for this experiment. Note that these test candidates have been excluded from the training samples. Each subject includes discrete images taken from different illumination conditions, facial expressions and poses. The proposed method was compared with the conventional AAMs [13] and its recent extension [19] proposed by Zhao et al.. The RMS point-point errors were measured against the ground truth annotation. The experimental results are presented in Figure 3(a). It shows that the proposed method yields more accurate performance than Zhao's method in most of the test cases. Specifically, in this particular dataset, by constraining shape and appearance variation at the same time, the proposed method yield an average 31% accuracy improvement than Zhao's method [19] which constrains appearance alone. To visualize the landmark registration performance, we have randomly selected test result from five test cases in Figure 3(b) 3(c). Both quantitative and qualitative results show that by constraining the shape and appearance in the ensemble, the proposed method produces more consistent landmark registrations for the discrete image ensemble.

IJAGS Database: The IJAGS database was collected in the earlier AAMs literature of [13]. Each sequence contains 180 frames. These videos were captured while the subjects were changing the head poses towards the camera and talking. In this experiment, we firstly sample the frames uniformly in time to select some key frames for our proposed method. The residue frames were then registered by the standard AAM fitting algorithm using the ensemble-specific AAMs determined from the key frames. The registration performances of the proposed method with different sample sizes were compared with the conventional generic AAMs, Zhao's method [19] and CLMs [15]. The Cumulative Distribution of the RMS registration errors of each sequence are presented in Figure 4. More detailed experimental results are demonstrated in Table 1. In this table, Da and Ds stand for the dimensionality of the ensemble-specific appearance and shape models determined by the proposed method. The original values of Da and Ds were defined in the generic AAM, which are 295 and 20. Since the frame numbers are much fewer than the appearance subspace dimension, then the original values of Da equals to the number of samples. The experimental results show that, (i) the proposed method outperforms the earlier work of generic AAMs, CLMs and [19] in terms of accuracy (on averages of 64.3%, 38.9% and 37.4% improvement respectively); (ii) the proposed method is able to determine a much lower dimension ensemble-specific model from a subset of sample frames, and this ensemble-specific model can be applied to the unsampled frames using the conventional AAMs fitting; (iii) the sampling strategy does not degrade the registration performance while saving significant computational time.

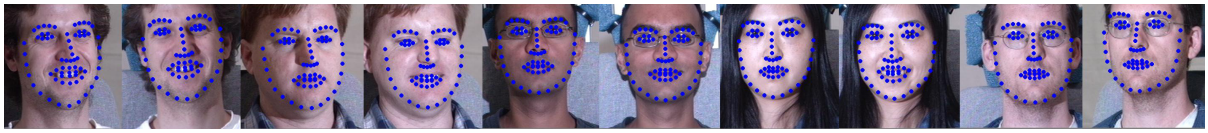
LFW Database: Labelled Faces in the Wild (LFW) [8] database is a collection of face photos taken under "real-life" conditions. It includes multiple images of the same subject with challenging poses, facial expressions, illumination conditions and some partial occlusions. In this experiment we employed a subset of 20 LFW subjects which



(a) The RMS Registration Errors



(b) Existing Method by Zhao et al. [19]



(c) The Proposed Method

Figure 3. (a): The RMS registration errors evaluated on the first 20 MultiPIE subjects, compared with existing method [19] and the conventional AAMs; (b): The landmark points registered by the existing method [19] on MultiPIE subject 1,6,9,16 and 18; (c): The landmark points registered by the proposed method.

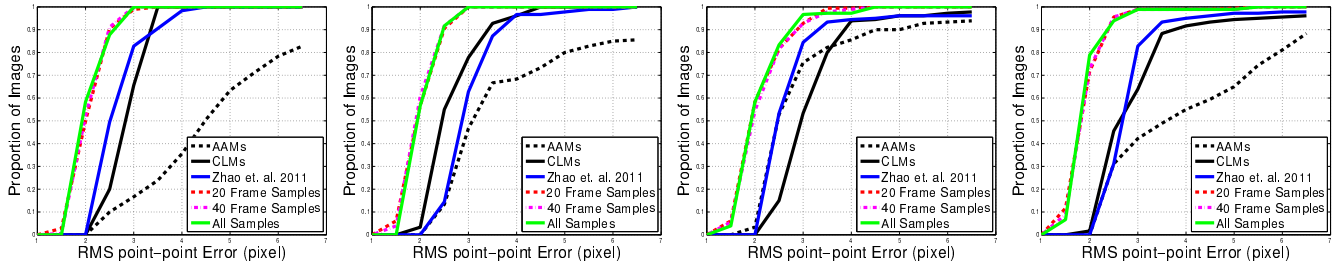
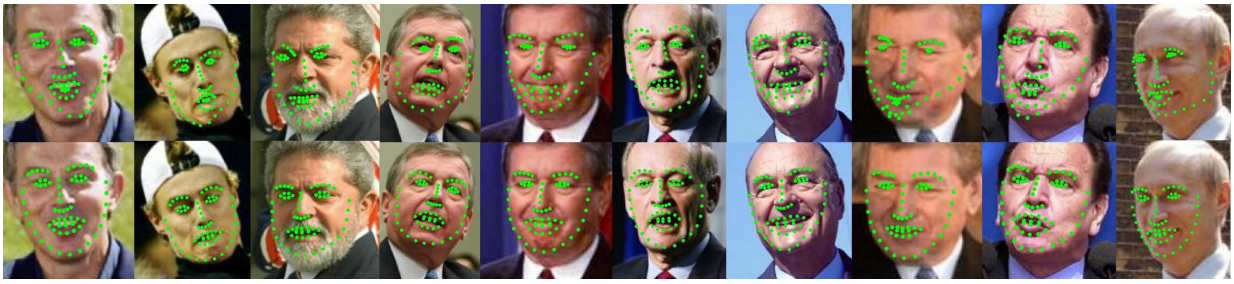


Figure 4. The Cumulative Distribution of the RMS registration errors evaluated on four subjects in IJAGS database. The proposed method with different sample sizes were compared with the existing method [19], conventional AAMs and CLMs.

was published together with the source code of [14]. In this experiment, we selected all the challenging photos included in LFW for each subject, then demonstrate a qualitative comparison between Zhao et al.'s [19] method and the proposed method with these challenging data. The alignment results in Figure 5 include four challenging cases, which are big facial expression variation, extreme lighting conditions, partial occlusion of face and image degration. The experimental results show that the proposed method produces impressive registration performance in these challenging test cases.

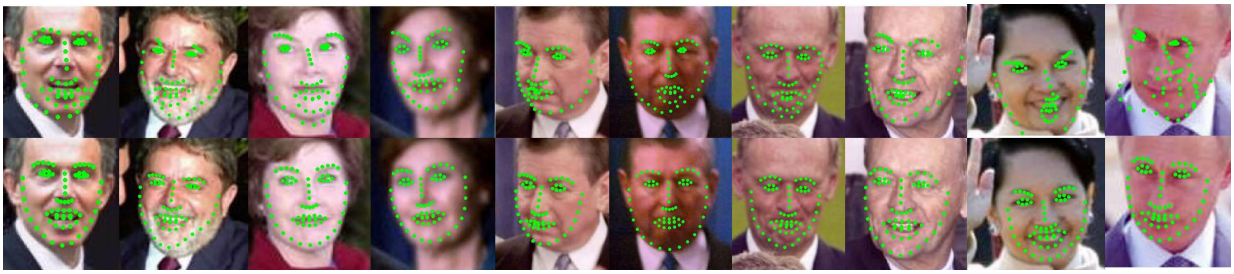
YouTube Celebrities Database: The proposed method was also evaluated on the YouTube Celebrities Face Tracking and Recognition Dataset [10]. This dataset was collected from the internet. It contains some "real-life" video clips. The registration result of three clips are demonstrated with the registered facial landmarks in Figure 6. The qualitative result shows that the proposed method is able to produce consistent registration performance on "real-life" videos with varying image conditions.

(a) Facial Expression



(b) Lighting

(c) Partial Occlusion



(d) Blurring

Figure 5. The registration performance of the existing method [19] (upper rows) and the proposed method (lower rows) tested on the images of the LFW database. The proposed method produces impressive registration performance on images with challenging conditions (big facial expression, large illumination variation, partial occlusion and image blurring) compared with the existing method.

7. Conclusion

In this paper, we propose a RASL inspired generic AAM fitting algorithm for image ensembles. By introducing rank constraints on both the generic appearance and shape subspaces, the proposed method is able to fit a generic AAM to unseen objects by automatically estimating the appropriate ensemble-specific AAM from the generic one. The proposed method advances earlier methods in three ways: (i) applying appearance and shape consistency instead of applying appearance alone produces more consistent alignments; (ii) using robust $||\cdot||_1$ norm on the facial appearance to improve the robustness to partial occlusions; and (iii) being able to determine the low dimension ensemble-specific AAM for additional images of same subject using standard AAMs fitting algorithms. Impressive experimen-

Subject	Samples	Da	Ds	Err	Time (mm:ss)
German	20	8	13	3.11	03:44
German	40	10	13	3.08	10:35
German	180	19	13	3.15	46:58
Simon	20	9	12	2.94	04:55
Simon	40	13	13	2.92	10:16
Simon	180	23	13	2.99	47:05
Jing	20	10	11	3.15	03:47
Jing	40	13	13	3.21	10:40
Jing	180	21	13	3.14	35:31
Iain	20	9	12	2.73	04:47
Iain	40	12	13	2.72	10:18
Iain	180	21	13	2.75	47:51

Table 1. Experimental Result of the proposed method applied on the IJAGS database with samples sizes of 20, 40 and all frames.

Figure 6. The registration performance on three video sequences of the YouTube Celebrities database. The proposed method produces consistent registration performance on video with complex background, bad resolution and big facial expressions.

tal results were demonstrated with a variety of challenging images and videos databases. Quantitative results show that the proposed method offers up to 37.4% improvement in the fitting accuracy compares with the state-of-the-art method.

8. Acknowledgement

This research was supported by an Australian Research Council (ARC) Discovery Research Grant DP110100827.

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. 2011.

[2] X. Cheng, S. Sridharan, J. Saragih, and S. Lucey. Anchored deformable face ensemble alignment. In *Computer Vision – ECCV 2012. Workshops and Demonstrations*, Lecture Notes in Computer Science, pages 133–142. 2012.

[3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *Computer Vision — ECCV'98*, Lecture Notes in Computer Science, pages 484–498. 1998.

[4] M. Cox, S. Lucey, S. Sridharan, and J. Cohn. Least squares congealing for unsupervised alignment of images. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[5] M. Cox, S. Sridharan, S. Lucey, and J. Cohn. Least-squares congealing for large numbers of images. In *Computer Vision, 2009 IEEE 12th International Conference on*, 2009.

[6] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(11):1080–1093, November 2005.

[7] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 2009.

[8] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[9] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53, 2000.

[10] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. *IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[11] E. G. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2006.

[12] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision (darpa). In *Proceedings of the 1981 DARPA Image Understanding Workshop*, pages 121–130, April 1981.

[13] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision (IJCV)*, 60, November 2004.

[14] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 763 – 770, 2010.

[15] J. M. Saragih, S. Lucey, and J. Cohn. Face alignment through subspace constrained mean-shifts. In *International Conference of Computer Vision (ICCV)*, September 2009.

[16] B. Smith and L. Zhang. Joint face alignment with non-parametric shape models. In *Computer Vision – ECCV 2012*, Lecture Notes in Computer Science. 2012.

[17] A. Vedaldi, G. Guidi, and S. Soatto. Joint data alignment up to (lossy) transformations. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008.

[18] J. Wright, Y. Ma, A. Ganesh, and S. Rao. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Proceedings of Neural Information Processing Systems (NIPS)*, 2009.

[19] C. Zhao, W.-K. Cham, and X. Wang. Joint face alignment with a generic deformable face model. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, 2011.