Deformable Face Ensemble Alignment with Robust Grouped-$\mathcal{L}1$ Anchors

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

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

***Abstract*— Many methods exist at the moment for deformable face fitting. A drawback to nearly all these approaches is that they are (i) noisy in terms of landmark positions, and (ii) the noise is biased across frames (i.e. the misalignment is toward common directions across all frames). In this paper we propose a grouped $\mathcal{L}1$-norm anchored method for simultaneously aligning an ensemble of deformable face images stemming from the same subject, given noisy heterogeneous landmark estimates. Impressive alignment performance improvement and refinement is obtained using very weak initialization as "anchors".**

I. Introduction

Deformable face fitting is the task of registering a parametrized face shape model to an image such that the points in the model (referred to as landmarks) correspond to consistent locations of interest (e.g. eye corners, mouth contour, etc.). 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 such as Active Appearance Models (AAMs) [11], Active Shape Models (ASMs) [3] and Constrained Local Models (CLMs) [14] have been proposed with varying degrees of success; however, these approaches often yield imperfect/noisy estimates of landmark positions.

Of particular interest to this paper is the task of simultaneously fitting a deformable face shape across an ensemble of images from very coarse initial alignments, and producing refined facial alignments. This task is closely related to the problem of unsupervised image ensemble alignment [9], [4], [12]. Recently, an approach referred to as Robust Alignment by Sparse and Low-rank (RASL) decomposition was proposed by Peng et al. [12]. 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 as there is nothing constraining the relative alignment. Without such constraint, the ensemble of face images can be considered aligned to a similar geometric frame of reference without looking like a face (see Figure 1). Recently, Zhao et. al. [16] proposed to constrain the RASL objective using a general facial appearance model. However, it has been well noted in AAM literature (Gross et. al. [5]) that it is hard to estimate a facial appearance basis that encompasses all possible human facial appearance variations (e.g. identity, lighting, age, expression, etc.). One of our central contributions in this paper is instead of constraining appearance, we employ only a constraint on the shape which is not affected by the appearance variations.

Fig. 1. (a) shows the registration errors of the CLMs tracker; (b) shows the face images transformed from the registered landmarks to the reference shape by piece-wise-affine [14]; (c) shows the transformed face images after attempting to refine the alignment using RASL method without "anchors"; (d) shows the low rank faces reconstructed from the refined alignment using the proposed method (grouped-$\mathcal{L}1$ anchor was used).

Specifically, we make four contributions in this paper. First, we study the error distribution of the state-of-the-art deformable face fitting methods and show that the errors are biased distributed. Second, we propose an anchoring strategy, grouped-$\mathcal{L}1$-norm, and demonstrate its ability to detect and "turn off" outlier anchor points automatically while only considering the constraints from inlier anchor points. Third, we propose a subsampled Anchored RASL algorithm for the alignment of an image ensemble with a large number of frames to improve the efficiency. Finally, we demonstrate state-of-the-art performance for deformable face fitting.

A. Notation

In this paper, we use consistent mathematical notations. Vectors are always presented in lower-case bold (e.g., $\mathbf{a}$). Matrices are in upper-case bold (e.g., $\mathbf{A}$) and scalars in lower-case (e.g. a). Images are expressed in capitalized form A. Warp functions $\mathcal{W}(\mathbf{x}_i;\mathbf{p}) = [\mathcal{W}_x(\mathbf{x}_i;\mathbf{p}), \mathcal{W}_y(\mathbf{x}_i;\mathbf{p})]^T$ will be used throughout this paper to denote a warping of the ith $2D$ coordinate vector $\mathbf{x}_i = [x_i, y_i]^T$ by a warp parameter vector $\mathbf{p} \in \mathcal{R}^P$, where P is the number of warp parameters, back to the ith position in a fixed base coordinate system. The concatenated vector of all discrete positions in the base coordinate system shall be defined as $\mathbf{z} = [x_1, \ldots, x_D, y_1, \ldots, y_D]^T$, similarly the warp function across all the concatenated coordinates shall be described as $\mathcal{W}(\mathbf{x};\mathbf{p}) = [\mathcal{W}_x(\mathbf{x}_1;\mathbf{p}), \ldots, \mathcal{W}_x(\mathbf{x}_D;\mathbf{p}), \mathcal{W}_y(\mathbf{x}_1;\mathbf{p}), \ldots, \mathcal{W}_y(\mathbf{x}_D;\mathbf{p})]^T$. 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 A is warped by the warp parameter vector $\mathbf{p}$, such that $A(\mathbf{p}) = [A(\mathcal{W}(\mathbf{x}_1;\mathbf{p})), \ldots, A(\mathcal{W}(\mathbf{x}_N;\mathbf{p}))]^T$. In this instance $A(\mathbf{p})$ is a N dimensional vector of image intensities, which N is substantially larger than D, since D represents the number of landmarks, and N stands for the number of pixels in the warped image. The steepest descent matrix $\frac{\partial A(\mathbf{p})}{\partial \mathbf{p}}$ of an image $A(\mathbf{p})$ is used frequently through out this paper. This $P \times N$ matrix is formed by combining image gradients of $A(\mathbf{p})$ 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 [11].

II. Robust Alignment by Sparse and Low-rank Decomposition (RASL)

RASL has become a popular method due to its robustness to illumination condition and appearance outliers (i.e. occlusion, disappearance/appearance of pixels) [12]. It is essentially a specific application of an earlier work called Robust Principal Component Analysis [15]. The central idea is to decompose the warped image ensemble matrix $\mathbf{D}(\mathbf{q}) = [(I_1(\mathbf{p}_1), \ldots I_F(\mathbf{p})]$ into a low rank appearance matrix $\mathbf{A}$ and a sparse errors matrix $\mathbf{E}$,

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

where $\mathbf{q} = [\mathbf{p}_1^T \ldots \mathbf{p}_F^T]^T$ is the super vector of warp parameters for all F frames in the image ensemble. The authors in [12] relaxed the objective by replacing $\text{rank}(\cdot)$ 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,

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

Note, since the relationship between the warp parameters $\mathbf{p}$ and the matrix of intensities $\mathbf{D}(\mathbf{p})$ is non-linear, a first order Taylor series linear approximation, $\mathbf{D}(\mathbf{q} + \Delta\mathbf{q}) \approx \mathbf{D}(\mathbf{q}) + \mathbf{J}(\mathbf{q})\Delta\mathbf{q}$, is employed, where $\mathbf{J}(\mathbf{q}) = [\frac{\partial I_1(\mathbf{p}_1)}{\partial \mathbf{p}_1}^T, \ldots, \frac{\partial I_F(\mathbf{p}_F)}{\partial \mathbf{p}_F}^T]^T$. Similar approximations are used within classical vision algorithms such as Lucas & Kanade (LK) image alignment [10], and Active Appearance Model fitting [11], but have the drawback of requiring an iterative solution to the objective, where $\mathbf{q} \leftarrow \mathbf{q} + \Delta\mathbf{q}$ is refined every iteration until convergence. Hitherto, RASL has been largely applied to simple linear warps such as affine and similarity. In our work we shall be employing more complex learned warps such as the Point Distribution Models (PDMs). Fortunately, like the canonical LK algorithm we have found the RASL algorithm to be largely warp agnostic as long as the non-linear nature of the warp is modeled appropriately.

III. Anchoring

In this section we present the modified RASL objective, which incorporates an additional penalty term to constrain landmark displacements,

$$\begin{aligned} \arg\min_{\mathbf{A},\mathbf{E},\boldsymbol{\Delta}\mathbf{q}} \quad & ||\mathbf{A}||_* + \lambda_1||\mathbf{E}||_1 + \lambda_2 \cdot \eta\{\mathbf{s}(\mathbf{q}) + \boldsymbol{\Phi}(\mathbf{q})\Delta\mathbf{q} - \bar{\mathbf{s}}\} \\ \text{s.t.} \quad & \mathbf{D}(\mathbf{q}) + \mathbf{J}(\mathbf{q})\Delta\mathbf{q} = \mathbf{A} + \mathbf{E}. \end{aligned} \tag{3}$$

We define $\mathbf{s}(\mathbf{q}) = [\mathcal{W}(\mathbf{x};\mathbf{p}_1)^T, \ldots, \mathcal{W}(\mathbf{x};\mathbf{p}_F)^T]^T$ as the super vector of F concatenated warped landmark vectors relating to each image in the ensemble, and $\boldsymbol{\Phi}(\mathbf{q}) = [\frac{\partial \mathcal{W}(\mathbf{x};\mathbf{p}_1)}{\partial \mathbf{p}_1}^T, \ldots, \frac{\partial \mathcal{W}(\mathbf{x};\mathbf{p}_F)}{\partial \mathbf{p}_F}^T]^T$ is the respective warp Jacobian matrix. The super vector $\bar{\mathbf{s}}$ are the concatenation of DF noisy landmark estimates, which we refer to herein as "anchors", and $\eta\{\}$ is the anchor penalty term. The central thesis of this paper is, without the anchor penalty, the canonical RASL objective will deform the subject's facial appearance in the image ensemble arbitrarily to find the minimum rank, in nearly all instances resulting in a false alignment.

A. Analysis on Alignment Errors of Face Trackers

A number of choices are available for the anchor penalty term $\eta\{\}$. In the earlier work of [2], the authors proposed an $\mathcal{L}2$-norm^2 function $\eta\{\mathbf{x}\} = ||\mathbf{x}||_2^2$, which implies a zero-mean noise assumption on the anchor landmark estimates $\mathbf{s}$. Unfortunately, as we will show in this section, alignment errors in commonly employed face fitting algorithms (e.g. CLMs) are not zero-mean. To study this, we utilized a public accessible face tracking toolbox FaceTracker [13] (an implementation of the well known CLMs algorithm [14]), to track face video sequences and determine the geometric errors by comparing the tracked landmarks to the manual annotated ground truth. Results from this analysis can be seen in Figure 2. From Figure 2(b) we can observe that some particular landmark points always misalign in the same direction across all frames in the sequence. For visual inspection on these points, in Figure 2(a), we randomly selected three frames from the sequence. We can visually observe that, in this particular video sequence, the landmark points on left hand side of jaw are always lower than the landmark points on the right hand side. More examples can be found in Figures 6(a) and 6(d).

From the results shown in Figure 2(b) and Figure 2(c) it can be observed that the biased errors (non-zero-mean) have two properties: (i) they are sparse, only a small subset of landmark points are misaligned, and (ii) they are typically larger than the normally distributed noise errors. Based on this analysis we argue that the use of an $\mathcal{L}2$-norm^2 anchoring term is not the best option, as it is well known that the performance of an objective function with an $\mathcal{L}2$-norm^2 constraint can be easily affected by the biased outliers [7].

IV. Deformable Face Ensemble Alignment with Grouped-$\mathcal{L}1$ Anchors

In this section, we introduce an alternate anchor penalty term. Instead of using $\mathcal{L}2$-norm^2 penalty, we propose to use the grouped-$\mathcal{L}1$-norm,

$$\eta\{\mathbf{x}\} = \sum_{f=1}^{F}\sum_{d=1}^{D} ||\mathbf{x}_{f,d}||_2{}^1 \tag{4}$$

where $\mathbf{x}_{f,d} = [x_{f,d}, y_{f,d}]^T$ is the 2D vector of $x-$ and $y-$ positions for the f-th frame and d-th landmark point. The motivation of using $\mathcal{L}1$-norm is that it is robust to outliers. $\mathcal{L}1$-norm is able to automatically select and "turn off" the outlier anchors while only considering the constraints from the inlier anchor points. For better definition of the alignment errors, for each point, we group $x_{f,d}$ and $y_{f,d}$ together to find the Euclidean distance of the misalignment $||\mathbf{x}_{f,d}||_2 = \sqrt{x_{f,d}^2 + y_{f,d}^2}$.

We can now re-write our original objective in Equation 3 with this new grouped-$\mathcal{L}1$ anchor term, an additional auxiliary variable $\mathbf{z}$, and the equality constraints,

$$\begin{aligned} \arg\min_{\mathbf{A},\mathbf{E},\Delta\mathbf{q},\mathbf{z}} \quad & ||\mathbf{A}||_* + \lambda_1||\mathbf{E}||_1 + \lambda_2 \cdot \sum_{f=1}^{F}\sum_{d=1}^{D} ||\mathbf{z}_{f,d}||_2 \\ \text{s.t.} \quad & \mathbf{D}(\mathbf{q}) + \mathbf{J}(\mathbf{q})\Delta\mathbf{q} = \mathbf{A} + \mathbf{E} \\ & \mathbf{z} = \mathbf{s}(\mathbf{q}) + \mathbf{\Phi}(\mathbf{q})\Delta\mathbf{q} - \bar{\mathbf{s}} \ . \end{aligned} \tag{5}$$

The introduction of the auxiliary variable $\mathbf{z}$ allows us to solve the objective efficiently using the Alternating Direction Method of Multipliers (ADMM) method [1]. The augmented Lagrangian can be written in scaled form [1] as,

$$\begin{aligned} &\mathcal{L}(\mathbf{A},\mathbf{E},\Delta\mathbf{q},\mathbf{z},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) = ||\mathbf{A}||_* + \lambda_1||\mathbf{E}||_1 + \\ &\lambda_2 \cdot \sum_{f=1}^{F}\sum_{d=1}^{D} ||\mathbf{z}_{f,d}||_2 + \frac{\mu}{2}||\mathbf{D}(\mathbf{q}) + \mathbf{J}(\mathbf{q})\Delta\mathbf{q} - \mathbf{A} - \mathbf{E} + \frac{1}{\mu}\boldsymbol{\xi}_1||_2^2 \\ &+ \frac{v}{2}||\mathbf{s}(\mathbf{q}) + \mathbf{\Phi}(\mathbf{q})\Delta\mathbf{q} - \bar{\mathbf{s}} - \mathbf{z} + \frac{1}{v}\boldsymbol{\xi}_2||_2^2, \end{aligned} \tag{6}$$

where $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are Lagrangian multipliers vectors, μ and v are positive scalars. The values of $\mathbf{A}$, $\mathbf{E}$, $\Delta\mathbf{q}$, $\mathbf{z}$ can be determined through a Gauss-Seidel style alternation strategy.

[1]We should note that, even though the $||\cdot||_1$ penalty does not appear in the objective, it is still considered as $\mathcal{L}1$ because it is the sum of the absolute Euclidean distances of the errors.

For every iteration k the parameters are updated as,

$$\mathbf{A}^{k+1} = \arg\min_{\mathbf{A}} \mathcal{L}(\mathbf{A}, \mathbf{E}^k, \mathbf{z}^k, \Delta\mathbf{q}^k, \boldsymbol{\xi}_1^k), \tag{7}$$

$$\mathbf{E}^{k+1} = \arg\min_{\mathbf{E}} \mathcal{L}(\mathbf{A}^{k+1}, \mathbf{E}, \mathbf{z}^k, \Delta\mathbf{q}^k, \boldsymbol{\xi}_1^k), \tag{8}$$

$$\mathbf{z}^{k+1} = \arg\min_{\mathbf{z}} \mathcal{L}(\mathbf{A}^{k+1}, \mathbf{E}^{k+1}, \mathbf{z}, \Delta\mathbf{q}^k, \boldsymbol{\xi}_2^k), \tag{9}$$

$$\Delta\mathbf{q}^{k+1} = \arg\min_{\Delta\mathbf{q}} \mathcal{L}(\mathbf{A}^{k+1}, \mathbf{E}^{k+1}, \mathbf{z}^{k+1}, \Delta\mathbf{q}, \boldsymbol{\xi}_1^k, \boldsymbol{\xi}_2^k), \tag{10}$$

$$\boldsymbol{\xi}_1^{k+1} = \boldsymbol{\xi}_1^k + \mu(\mathbf{D}(\mathbf{q}) + \mathbf{J}(\mathbf{q})\Delta\mathbf{q}^{k+1} - \mathbf{A}^{k+1} - \mathbf{E}^{k+1}), \tag{11}$$

$$\boldsymbol{\xi}_2^{k+1} = \boldsymbol{\xi}_2^k + v(\mathbf{s}(\mathbf{q}) + \mathbf{\Phi}(\mathbf{q})\Delta\mathbf{q}^{k+1} - \bar{\mathbf{s}} - \mathbf{z}^{k+1}), \tag{12}$$

$$\mu^{k+1} = a \cdot \mu^k, \tag{13}$$

$$v^{k+1} = a \cdot v^k. \tag{14}$$

Here a is an incremental factor for the scalars μ and v. In our implementation, the most efficiency was found using $a = 1.25$ by experiments.

A. Efficient Sub-Problems

ADMMs are extremely efficient as they enable one to break a complex objective into a sequence of efficient sub-problems. The sub-problems 7, 8 and 9 can be solved efficiently by the soft threshold methods [12], [1],

$$(\mathbf{U}, \mathbf{\Sigma}, V) = \text{svd}(\mathbf{D}(\mathbf{q}) + \mathbf{J}(\mathbf{q})\Delta\mathbf{q}^k - \mathbf{E}^k + \frac{1}{\mu}\boldsymbol{\xi}_1^k), \tag{15}$$

$$\mathbf{A}^{k+1} = \mathbf{U}\mathcal{S}_{\frac{1}{\mu}}[\mathbf{\Sigma}]\mathbf{V}^{\mathrm{T}}, \tag{16}$$

$$\mathbf{E}^{k+1} = \mathcal{S}_{\frac{\lambda_1}{\mu}}[\mathbf{D}(\mathbf{q}) + \mathbf{J}(\mathbf{q})\Delta\mathbf{q}^k - \mathbf{A}^{k+1} + \frac{1}{\mu}\boldsymbol{\xi}_1^k], \tag{17}$$

$$\mathbf{z}_{f,d}^{k+1} = \mathcal{T}_{\frac{\lambda_2}{v}}[\mathbf{s}(\mathbf{q}) + \mathbf{\Phi}(\mathbf{q})\Delta\mathbf{q}^k - \bar{\mathbf{s}} + \frac{1}{v}\boldsymbol{\xi}_2^k]_{i,j}, \tag{18}$$

where the soft threshold operators $\mathcal{S}$ and $\mathcal{T}$ are defined as:

$$\mathcal{S}_k[a] = \begin{cases} a - k & \text{if } a > k, \\ 0 & \text{if } |a| <= k, \\ a + k & \text{if } a < -k, \end{cases} \tag{19}$$

or equivalently,

$$\mathcal{S}_k[a] = (a - k)_+ - (-a - k)_+, \tag{20}$$

and,

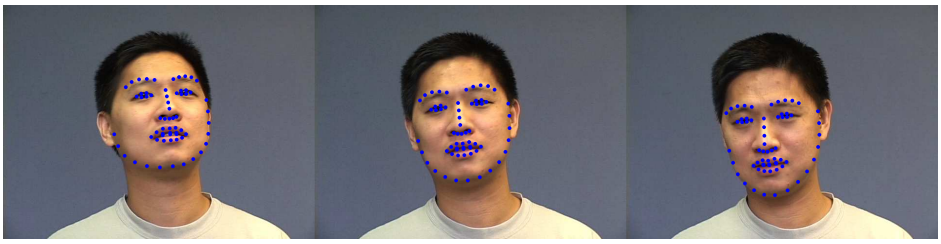
$$\mathcal{T}_k[a] = (1 - k/||a||_2)_+ a. \tag{21}$$

Equation 10 can be solved efficiently as a least squares problem,

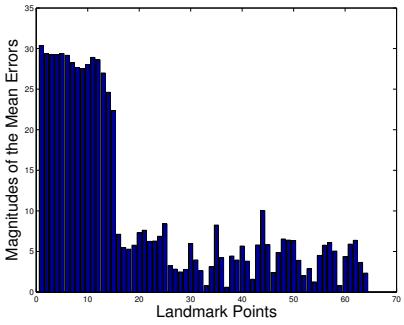
$$\begin{aligned} \Delta\mathbf{q}^{k+1} = \arg\min_{\Delta\mathbf{q}} \frac{\mu}{2}||\mathbf{D}(\mathbf{q}) + \mathbf{J}(\mathbf{q})\Delta\mathbf{q} - \mathbf{A}^{k+1} - \mathbf{E}^{k+1} + \\ \frac{1}{\mu}\boldsymbol{\xi}_1^k||_2^2 + \frac{v}{2}||\mathbf{s}(\mathbf{q}) + \mathbf{\Phi}(\mathbf{q})\Delta\mathbf{q} - \bar{\mathbf{s}} - \mathbf{z}^{k+1} + \frac{1}{v}\boldsymbol{\xi}_2^k||_2^2 \ . \end{aligned} \tag{22}$$

V. An Efficient Framework for Large Scale Alignments

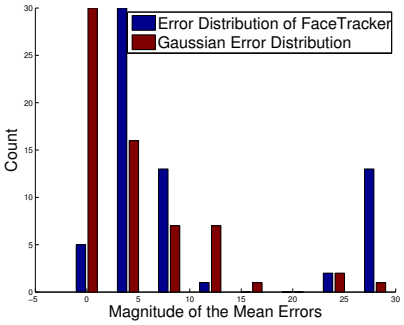
Through the employment of efficient optimization strategies like ADMMs RASL can be applied to reasonably large image ensembles. However, the computational cost of RASL is exponential as a function of the number of images in the ensemble, making the approach impractical for face sequences with tens of thousands of frames. In this section, we



(a)



(b)



(c)

Fig. 2. (a) shows three randomly selected frames of the IJAGS database aligned by CLMs [13]. It can be observed that for some particular points, the alignment errors are in same directions; (b) is the absolute magnitude of the mean errors for all points across all frames; (c) is the histogram of absolute magnitude of mean errors. It can be observed that the error distribution of a face tracker is not zero-mean.

propose an efficient framework which applies the anchored RASL method to a subset of key frames to find a low rank appearance basis $\mathbf{A}$. This basis can then be employed within a canonical person-specific Active Appearance Model fitting strategy to efficiently track the residual frames. This strategy is especially efficient as one can employ the "project-out" AAM fitting algorithm.

$$\arg\min_{\mathbf{p},\boldsymbol{\lambda}} ||I(\mathbf{p}) - A_0(\mathbf{0}) + \mathbf{A}\boldsymbol{\lambda}||^2, \tag{23}$$

where one solves simultaneously for the warp $\mathbf{p}$ and $\boldsymbol{\lambda}$ appearance parameters. A number of approaches have been proposed in the literature for fitting AAMs [11]. The most notable and popular of these variants are approaches based on the LK algorithm [10]. Like RASL, the objective function in Equation 23 is difficult to solve as their is a non-linear relationship between the shape parameters $\mathbf{p}$ and the appearance parameters $\boldsymbol{\lambda}$. A key insight, stemming from Lucas & Kanade [10], was that a linear approximation can be made between $\mathbf{p}$ and $\boldsymbol{\lambda}$ through the judicious use of image gradients and the chain rule to form steepest descent matrices (i.e. $\frac{\partial A(\mathbf{p})}{\partial \mathbf{p}}$). In particular the employment of inverse compositional extensions can allow for extremely efficient per-image fitting.

VI. Experiments

In this section we evaluate the performance of our Anchored RASL method on a variety of face alignment tasks. The PDMs emplyed in this paper was learnt from the landmark points of all subjects of the IJAGS [11] database and MultiPIE [6] database (5 subjects of IJAGS and 346 subjects in MultiPIE, with varying head poses and facial expressions). The obtained PDMs consists of 19 degrees of freedom with 66 landmark points. The image in the reference shape frame was scaled to 10,000 pixels in each of the Red, Green and Blue channels. The weight, λ_1 was selected using the same strategy as in [12], $\lambda_1 = 1/\sqrt{N}$, where N is the number of pixels in each aligned image (30,000 in our case). The experimental result shows that the best performance was found when using $\lambda_2 = 0.3/\sqrt{D}$, where D is the number of landmark points in every frame (66 in our implementation).

A. Simulation

To verify the robustness of our method to the biased errors in the anchor points, we conducted a simulation using synthetic data. A sequence of 40 frames of the same subject with large head pose variations were selected from the IJAGS database. The 66 landmark points of each frame were manually annotated as the ground truth. We randomly selected a sparse subset of N points. For the selected landmark points, synthetic error was added to the ground truth in every frame to create a biased/directional misalignment as anchors and the initial landmark estimates. Two experiments were conducted. Firstly we compared the performance of the $\mathcal{L}2$-norm^2 method, grouped-$\mathcal{L}1$-norm method and the $\mathcal{L}1$-norm method (without grouping) by selecting $N = 13$ or equivalently 20% of the anchor points as outliers and perturbed them with increasing error (Figure 3(a)). The experimental results show that in $\mathcal{L}1$-norm or grouped-$\mathcal{L}1$-norm methods, increasing the biased errors doesn't affect the final result. This is because the outliers have been successfully detected and ignored by the $\mathcal{L}1$ anchoring term.

In the second experiment, we compared the performance of the three anchored RASL methods by increasing the

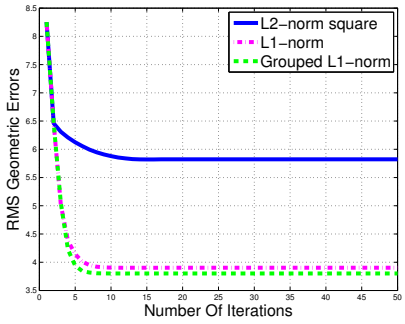


Fig. 4. The RMS geometric errors. It can be observed that the $\mathcal{L}1$-norm and the grouped-$\mathcal{L}1$-norm have better fitting performance and faster convergence rate.

number of outliers n (Figure 3(b)). We can observe that $\mathcal{L}1$-norm and grouped-$\mathcal{L}1$-norm perform much better than the $\mathcal{L}2$-norm^2 method, where the best performance was found by using grouped-$\mathcal{L}1$-norm. Furthermore, the result shows that $\mathcal{L}1$-norm and grouped-$\mathcal{L}1$-norm are able to detect outliers accurately in a sequence with up to 22% outliers.

B. Evaluation using Controlled Data

To evaluate our method with anchor points generated by the state-of-the-art face tracker, we initialized our Anchored RASL approach by a public CLMs face tracker implementation, FaceTracker [13], on a subset of video sequence of a single subject in the IJAGS database. Two experiments were conducted. In the first experiment we aimed to study the convergence performances of the three different anchoring strategies. In every iteration, the RMS geometric errors were determined by comparing the current landmark estimate and the ground truth in the reference shape frame. The experiment result (Figure 4) shows that $\mathcal{L}1$-norm and grouped-$\mathcal{L}1$-norm methods significantly outperform the $\mathcal{L}2$-norm^2 method in terms of the alignment performance and the rate of convergence.

In the second experiment, we have conducted an evaluation on the proposed efficient alignment algorithm. In this experiment, computational time and the alignment performance have been evaluated at different numbers of RASL sub-samples. The experimental result (Figure 5) shows that this algorithm is able to reduce computational cost significantly while maintaining good alignment performance (i.e. At 180 frames, 84% of computational cost was saved by subsampling 20 frames for the Anchored RASL).

C. Experimental Comparison with Existing method

To compare the performance of our method with Zhao et. al's approach [16], we have conducted two experiments. In the first experiment, we selected 40 frames of each IJAGS subject for evaluation. In our implementation, the generic facial appearance model used in Zhao et. al.'s method was obtained from all 346 subjects of MultiPIE database. This came with 181 appearance basis. Both methods are initialized with CLMs, and the final geometric errors are determined by the RMS point-point error in the reference shape. The experimental results are presented in Table I.

	Adrian	German	Iain	Jing	Simon
Zhao et. al. [16]	9.15	7.53	8.34	8.27	5.31
Anchored RASL	4.31	3.84	3.70	5.35	3.63

TABLE I
EXPERIMENTAL COMPARISON OF THE RMS GEOMETRIC ERRORS BETWEEN ZHAO'S METHOD AND OUR METHOD WITH IJAGS DATABASE

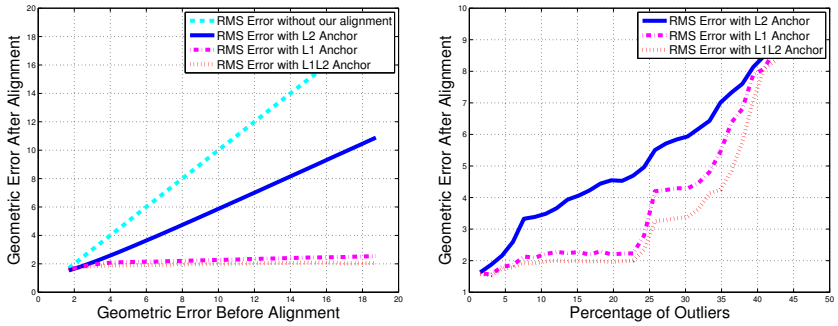
To visually compare the performance of each method with image sequences taken under uncontrolled conditions, in the second experiment we applied both alignment approaches to two video clips of television interviews collected by the YouTube Celebrities Face Tracking and Recognition Dataset [8]. For the visualization of the alignment results, selected frames of each sequence are demonstrated with the plotted landmark positions of the initial alignments by CLMs (Saragih et. al. [14]) and the final refined alignments using each method. The experimental results show that our method significantly outperforms the earlier method in terms of alignment accuracy, even when the initial alignment is very noisy and the quality of data is very poor (low resolution, complex background etc.). Although the result in Figure 6(f) is not as perfect as we expected, it has still refined the landmarks to a reasonable location from the extreme initialization as shown in Figure 6(d).

VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed an alternative RASL extension approach for deformable face ensemble alignment. We employed a constraint on the shape of the face to enable RASL for deformable face alignment tasks. To choose the best way to apply the shape constraint, we analysed the tracking error of the state-of-the-art face tracker, and showed that most alignment errors are biased (the mean errors are not zero). Then we proposed to constrain the RASL objective with grouped-$\mathcal{L}1$-norm anchoring and showed that grouped-$\mathcal{L}1$-norm method is more robust to the biased outliers, compared with the $\mathcal{L}2$-norm^2 method. To enable our method for large video sequences, we proposed an efficient framework which significantly reduces the computational cost by using RASL and AAM serially on different subsets of frames. Experimental results show that our method outperformed the earlier method in terms of the alignment accuracy. The visual results show the proposed method is able to refine the alignment even when the data is of poor quality (low resolution, complex background, and big facial expressions).

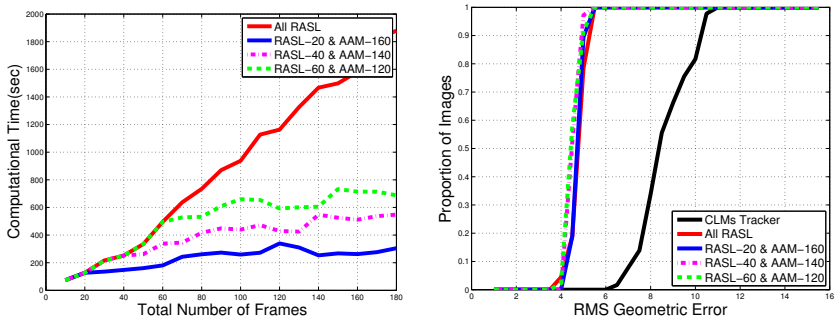
In this paper the grouped-$\mathcal{L}1$ constraint was proposed as we believed points should be treated as a whole (not separated into x- and y- components). One of the unexpected results of this paper is that it seems to make little difference in practice. We shall explore this phenomenon further in the future.

Acknowledgement: This work is supported by the Australian Research Council (ARC) Discovery Grant No: DP1110827.



(a) (b)

Fig. 3. Performance comparison between anchors of $\mathcal{L}2$-norm^2, $\mathcal{L}1$-norm and grouped-$\mathcal{L}1$-norm (referred to as $\mathcal{L}1\mathcal{L}2$-norm); (a) The fitting performances with increasing biased errors. (b) The fitting performances with increasing number of outliers.



(a) Computational Time (b) Fitting Performance Curve

Fig. 5. (a) The computational time of the sub-sample sizes of 180 (all), 20, 40, 60. (b) The fitting performance curve (the proportion of frames versus the maximum RMS geometric error). The experimental results show that our algorithm is able to reduce computational cost significantly while maintaining the identical alignment performance

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*, volume 3. 2011.

[2] X. Cheng, S. Sridharan, J. Saragih, and S. Lucey. Anchored deformable face ensemble alignment. In A. Fusiello, V. Murino, and R. Cucchiara, editors, *Computer Vision ECCV 2012. Workshops and Demonstrations*, volume 7583 of *Lecture Notes in Computer Science*, pages 133–142. Springer Berlin Heidelberg, 2012.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models their training and application. *Computer Vision and Image Understanding (CVIU)*, 61(1):38–59, Jan. 1995.

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

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

[7] Q. Ke and T. Kanade. Robust subspace computation using l1 norm. In *CMU Technical Report CMU-CS-03-172*, Auguest 2003.

[8] 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.

[9] E. G. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28:236–250, 2006.

[10] 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.

[11] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision (IJCV)*, 60(2):135 – 164, November 2004.

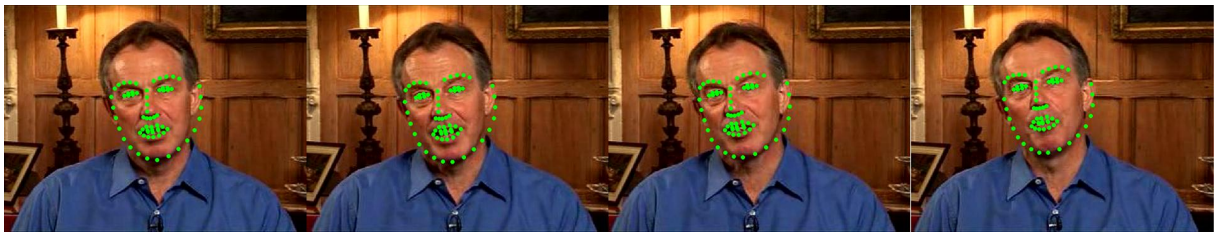
[12] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 763 –770, june 2010.

[13] J. Saraghi. Facetracker. In *http://web.mac.com/jsaragih/FaceTracker/FaceTracker.html*, 2011.

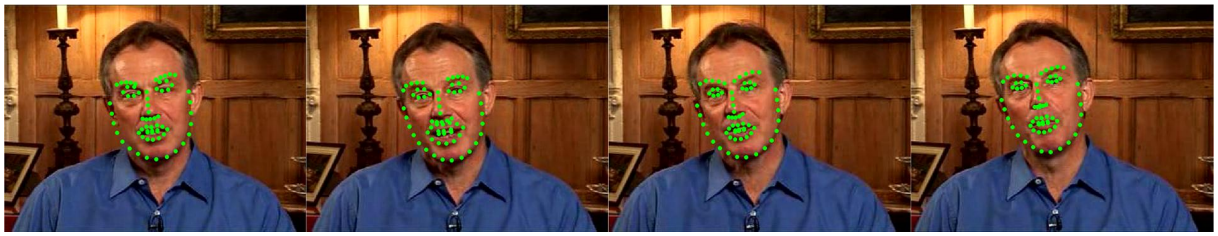
[14] J. M. Saragih, S. Lucey, and J. Cohn. Face alignment through subspace constrained mean-shifts. In *International Conference of Computer Vision (ICCV)*, September 2009.

[15] 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.

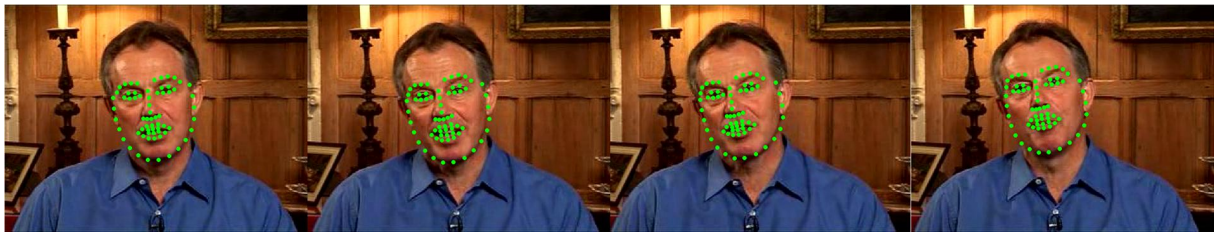
[16] C. Zhao, W.-K. Cham, and X. Wang. Joint face alignment with a generic deformable face model. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 561–568, Washington, DC, USA, 2011. IEEE Computer Society.



(a) Tony Blair - Saragih et. al. [14]



(b) Tony Blair - Zhao et. al. [16]



(c) Tony Blair - Our proposed Anchored RASL

(d) Bill Clinton - Saragih et. al. [14]

(e) Bill Clinton - Zhao et. al. [16]

(f) Bill Clinton - Our proposed Anchored RASL

Fig. 6. Alignment performances of the initial CLMs alignment, Zhao et. al.'s method and our grouped-$\mathcal{L}1$-norm Anchored RASL method. (a)(b)(c) Faces in complex background which have similar color as human skin. (d)(e)(f) Faces in extreme condition, low resolution, complex background, and big facial expressions. The alignments are refined by our method even when the data is of poor quality.