Elsevier Editorial System(tm) for Image and Vision Computing
Manuscript Draft

Manuscript Number:

Title: Non-Rigid Face Tracking with Enforced Convexity and Local Appearance Consistency Constraint

Article Type: Best of FG 2008

Keywords: Constrained Local Models, Convex Quadratic Fitting, Non-Rigid Face Tracking

Corresponding Author: Dr. Simon Lucey, Ph.D.

Corresponding Author's Institution: Carnegie Mellon University

First Author: Simon Lucey, Ph.D.

Order of Authors: Simon Lucey, Ph.D.; Yang Wang, Ph.D.; Jason Saragih, Ph.D. ; Jeffery F Cohn, Ph.D.

Abstract: Convex quadratic fitting (CQF) has demonstrated great success recently in the task of non-rigidly registering a face in a still image using a constrained local model (CLM). A CLM is a commonly used model for non-rigid object registration and contains two components: (i) local patch-experts that model the appearance of each landmark in the object, and (ii) a global shape prior describing how each of these landmarks can vary non-rigidly. Conventional CLMs can be used in non-rigid facial tracking applications through a track-by-detection strategy. However, the registration performance of such a strategy is susceptible to local appearance ambiguity. Since there is no motion continuity constraint between neighboring frames of the same sequence, the resultant object alignment might not be consistent from frame to frame and the motion field is not temporally smooth. In this paper, we extend the CQF fitting method into the spatio-temporal domain by enforcing the appearance consistency constraint of each local patch between neighboring frames. More importantly, we show, as in the original CQF formulation, that the global warp update can be optimized jointly in an efficient manner. Finally, we demonstrate that our approach receives

improved performance for the task of non-rigid facial motion tracking on the videos of clinical patients.

Title:- Non-Rigid Face Tracking with Enforced Convexity and Local Appearance Consistency Constraint

Abstract:- Convex quadratic fitting (CQF) has demonstrated great success recently in the task of non-rigidly registering a face in a still image using a constrained local model (CLM). A CLM is a commonly used model for non-rigid object registration and contains two components: (i) local patch-experts that model the appearance of each landmark in the object, and (ii) a global shape prior describing how each of these landmarks can vary non-rigidly. Conventional CLMs can be used in non-rigid facial tracking applications through a track-by-detection strategy. However, the registration performance of such a strategy is susceptible to local appearance ambiguity. Since there is no motion continuity constraint between neighboring frames of the same sequence, the resultant object alignment might not be consistent from frame to frame and the motion field is not temporally smooth. In this paper, we extend the CQF fitting method into the spatio-temporal domain by enforcing the appearance consistency constraint of each local patch between neighboring frames. More importantly, we show, as in the original CQF formulation, that the global warp update can be optimized jointly in an efficient manner. Finally, we demonstrate that our approach receives improved performance for the task of non-rigid facial motion tracking on the videos of clinical patients.

Keywords:- Constrained Local Models, Convex Quadratic Fitting, Non-Rigid Face Tracking

Non-rigid Face Tracking with Enforced Convexity and Local Appearance Consistency Constraint

Simon Lucey,* Yang Wang, Jason Saragih, Jeffery F. Cohn

Robotics Institute, Carnegie Mellon University, Pittsburgh PA 15213, USA

Abstract

Convex quadratic fitting (CQF) has demonstrated great success recently in the task of non-rigidly registering a face in a still image using a constrained local model (CLM). A CLM is a commonly used model for non-rigid object registration and contains two components: (i) local patch-experts that model the appearance of each landmark in the object, and (ii) a global shape prior describing how each of these landmarks can vary non-rigidly. Conventional CLMs can be used in non-rigid facial tracking applications through a track-by-detection strategy. However, the registration performance of such a strategy is susceptible to local appearance ambiguity. Since there is no motion continuity constraint between neighboring frames of the same sequence, the resultant object alignment might not be consistent from frame to frame and the motion field is not temporally smooth. In this paper, we extend the CQF fitting method into the spatio-temporal domain by enforcing the appearance consistency constraint of each local patch between neighboring frames. More importantly, we show, as in the original CQF formulation, that the global warp update can be optimized jointly in an efficient manner. Finally, we demonstrate that our approach receives improved performance for the task of non-rigid facial motion tracking on the videos of clinical patients.

Key words: Constrained Local Models, Convex Quadratic Fitting, Non-Rigid Face Tracking

* Corresponding author.
Email addresses: slucey@cs.cmu.edu (Simon Lucey,), wangy@cs.cmu.edu (Yang Wang,), jsaragih@andrew.cmu.edu (Jason Saragih,), jeffcohn@cs.cmu.edu (Jeffery F. Cohn).

1 Introduction

Accurate and consistent tracking of non-rigid object motion, such as facial motion and expressions, is important in many computer vision applications and has been studied intensively in the last two decades [2,6,27,26,8,13,17,9,10,1,25,18]. This problem is particularly difficult when tracking subjects with previously unseen appearance variations. To address this problem, a number of registration/tracking methods have been developed based on local region descriptors and a non-rigid shape prior [8,13,14,17,18,22,23,8]. We refer to these family of methods collectively as a constrained local model (CLM) [1] Probably, the best known example of a CLM can be found in the seminal active shape model (ASM) work of Cootes and Taylor [7]. Instantiations of CLMs differ primarily in the literature with regards to: (i) whether the local experts employ a 1D or 2D local search, (ii) how the local experts are learnt, (iii) how the source image is normalized geometrically and photometrically before the application of the local experts, and (iv) how one fits the local experts responses to conform to the global non-rigid shape prior. Disregarding these differences, however, all instantiations of CLMs can be considered to be pursuing the same two goals: (i) perform an exhaustive local search for each landmark around their current estimate using some kind of patch-expert (i.e., feature detector), and (ii) optimize the global non-rigid shape parameters such that the local responses for all of its landmarks are minimized.

Compared to the holistic representations, such as active appearance models (AAMs), CLMs offer many advantages when registering "real-world" face images. First, the ability to employ photometric normalization at each local expert. Second, the ability to handle larger mismatches than gradient methods in initial registration due to the use of local search for each landmark. Third, in comparison to holistic AAMs, CLMs have inherent computational advantages and can be easily applied to parallel computation architectures. Finally, the ability to employ local experts that have been discriminatively trained from large hand labeled offline face datasets and exhibit good generalization performance on unseen images. This is in contrast to holistic AAM methods that have demonstrated poor generalization when learning from very large datasets [11].

CLMs can be equally applied to tracking applications as single-image registration applications through the employment of a track-by-detection paradigm. In this approach the CLM is initialized by the result of the preceding frame in the image sequence where on then re-applies the fitting process. However,

[1] Our definition of CLMs is much broader than that given by Cristinacce and Cootes [8] who employ the same name for their approach. Cristinacce and Cootes' method can be thought of as a specific subset of the CLM family of models.

the matching performance of the local patch-experts might be susceptible to local appearance ambiguity. Since there is no motion continuity constraint between neighboring frames of the same sequence, the resultant alignment might not be consistent from frame to frame and the motion field is not temporally smooth.

Inspired by recent work for aligning a set of images in an unsupervised manner [3,16,15,22] we propose a new approach to achieve accurate and consistent tracking of non-rigid object motion in a video sequence by extending the CLM method into the spatio-temporal domain. By enforcing the appearance consistency constraint of each local patch between neighboring frames, the temporal texture coherence is integrated into the CLM framework as a motion smoothness constraint. We make the following contributions in our paper:

- We extend the constrained local model (CLM) method into the spatio-temporal domain by introducing the appearance consistency constraint of each local patch between neighboring frames. Furthermore, to incorporate this local appearance consistency constraint efficiently into the CLM framework, we compute the image error in different reference frames, i.e., between the input image and the model images from previous frames. (Section 3)
- We show that a specific form of the classic Lucas-Kanade [19] approach to gradient-descent image alignment can be viewed as a CLM where each local response surface is *indirectly* approximated through a convex quadratic function. Since each of the approximated response surfaces are convex an explicit solution to the approximate joint minima can be found (since it too is convex). This process can be iterated until some convergence towards the actual joint minima is obtained. (Section 4)
- Instead of using computationally expensive generic optimizers such as the Nelder-Mead simplex [8] method, we propose a *convex quadratic fitting* (CQF) approach that is able to *directly* fit a convex quadratic to both the local response surface of a local patch-expert and the associated local appearance consistency constraints. Since each of the approximated response surfaces is convex, an explicit solution to the approximate joint minima can be found. As a result, we are able to apply a similar optimization as employed in the Lucas-Kanade algorithm within the generic CLM framework. (Section 4 and 6)
- Finally, we demonstrate improved non-rigid face tracking performance on the video sequences in a clinical archive which contains video clips of pain patients. Our extended CLM approach exhibits superior performance to the CLM approach without the local appearance consistency constraint and leading holistic AAM [6] approaches to non-rigid object tracking. (Section 7)

2 Learning Constrained Local Models

The notation employed in this paper shall depart slightly from canonical methods in order to easily allow the inclusion of patches of intensity at each coordinate rather than just pixels. When a template T is indexed by the coordinate vector $\mathbf{x} = [x, y]^T$ it not only refers to the pixel intensity at that position, but the local support region (patch) around that position. For additional robustness the $P \times P$ support region [2] is extracted after the image has been suitably normalized for scale and rotation to a base template of the non-rigid object. $T(\mathbf{x}_k)$ and $Y(\mathbf{x}_k)$ refer to the vector concatenation of image intensity values within the kth region (patch) of the template image T and the source image Y, respectively.

2.1 Estimating Patch Experts

The choice of classifier employed to learn patch experts within a CLM can be considered to be largely arbitrary allowing the use of generative (e.g., Gaussian likelihood function) or discriminative (e.g., support vector machine (SVM), AdaBoost, relevance vector machines (RVM), etc.) local models. We chose to use a SVM in our work due to its ability to discriminatively learn a local expert as well as generalize from thousands training examples. A linear SVM was chosen in our work over other non-linear kernel varieties due to its computational advantages in that,

$$\begin{aligned}\hat{f}(\Delta\mathbf{x}) &= \sum_{i=1}^{N_S} \gamma_i \alpha_i T_i(\mathbf{x})^T Y(\mathbf{x} + \Delta\mathbf{x}) \\ &= Y(\mathbf{x} + \Delta\mathbf{x})^T \sum_{i=1}^{N_S} \gamma_i \alpha_i T_i(\mathbf{x}) \end{aligned} \tag{1}$$

where $\hat{f}(\Delta\mathbf{x})$ is the match-score for the patch-expert at coordinate displacement $\Delta\mathbf{x}$ from the current patch coordinate center $\mathbf{x}$. Y is the source image, T_i is the ith support vector, α_i is the corresponding support weight, $\gamma_i \in \{not\ aligned\ (-1), aligned\ (+1)\}$ is the corresponding support label, and N_S is the number of support vectors.

Training SVM classifiers: Employing a linear SVM is advantageous as it allows for $\sum_{i=1}^{N_S} \gamma_i \alpha_i T_i(\mathbf{x})$ to be pre-computed for Equation 1 rather than evaluated at every $\Delta\mathbf{x}$. The support images T_i are obtained from an offline

[2] A typical patch size is 15×15 in our experiments for a face object with an inter-ocular distance of 50 pixels.

training set of positive and negative images. Positive patch examples were obtained for patches centered at the fiduciary points of our training images, while negative examples were obtained by sampling patches shifted away from the ground truth. In our experiments a 15×15 window, centered at each ground truth position, is used to obtain the image patches for the positive examples. The negative examples are obtained by shifting the above window within 15 pixels to each ground truth position.

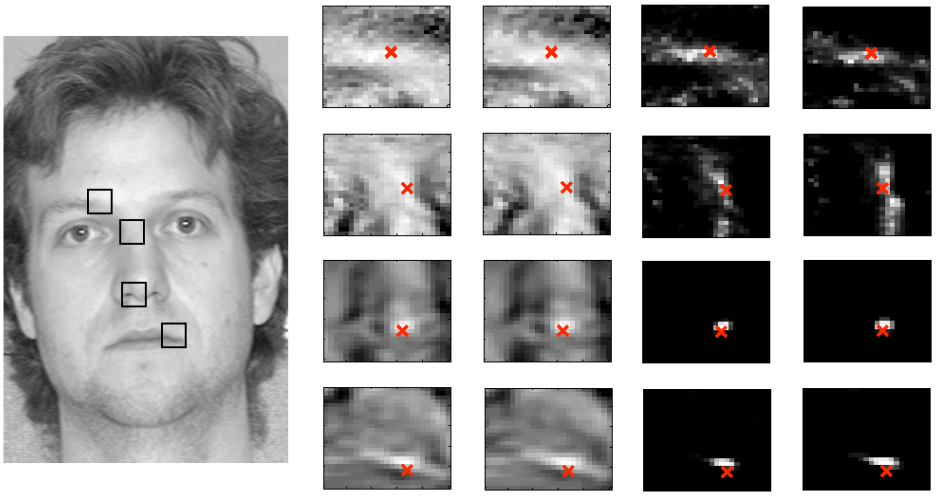
In order to get a good decision boundary in the SVM training, two sampling strategies are adopted to reduce the similarity between the positive and negative examples. First, we enforce the center of the negative sampling window to be at least 2 or 3 pixels away from the ground truth position. Second, for the fiduciary points on a contour (e.g., the points along the jaw line), it is reasonable to constrain the search only in the normal direction. Furthermore, to speed up the training process, we sort the negative examples based on the sum of squared differences (SSD) between the negative and positive examples in the descending order, and select the top $5-10\%$ negative examples for the SVM training. As demonstrated in Figure 1, the performance of the patch experts learned by a smaller training set, shown in Figure 1(c) and (e), is almost the same as the performance seen for experts trained on a larger number of training examples in Figure 1(b) and (d). A small subset (5%) of the original negative examples are used in our experiments.

Obtaining Local Responses: Once the patch expert has been trained we can obtain a local response for an individual patch expert by performing an exhaustive search of the neighboring region of that patch's current position within the source image. In our experiments, we found a search window size of 15×15 pixels for each patch gave good results for a face object with an inter-ocular distance of 50 pixels.

Example response surfaces are shown in Figure 1. To illustrate the effectiveness of our patch experts we placed the center of the searching window randomly away from the ground truth position. From the top row to the bottom in Figure 1(b-e), it shows the local responses for patch experts describing the left eyebrow, the nose bridge, the nose end, and the right mouth corner, respectively. As one can see, the estimated responses perform a good job of finding the ground truth location. All response surfaces were obtained from a linear SVM.

In Figure 1(b), 125 positive examples and $15k$ negative examples were used to train each patch expert, while in Figure 1(c), 125 positive examples and $8k$ negative examples were used. Both positive and negative examples contained 15×15 patches extracted from the training images. As we can see, the performance of the patch experts learned by a smaller training set, shown in Figure 1(c) and (e), is almost the same as the performance seen for ex-

perts trained on a larger number of training examples in Figure 1(b) and (d). This result demonstrated that our patch-experts had a reasonable amount of training examples for employment within a CLM framework.



(a) (b) (c) (d) (e)

Fig. 1. Examples of local search responses: (a) is the source image to be aligned. (b) shows the local search responses using patch experts trained by 125 positive examples and $15k$ negative examples. (c) shows the local search responses trained by 125 positive examples and $8k$ negative examples. (d) and (e) show the estimated logistic regression weight values of (b) and (c), respectively. A high intensity value indicates a small matching error between the template and the source image patch. Each row in (b-e) shows the responses and weights within a 25×25 local search window. The location of each search window is illustrated in the source image (a) as a black box, while the red cross illustrate the ground truth alignment. It is interesting to see that the patch experts learned by a smaller training set (including $8k$ negative examples) have very similar performance as the ones trained by large training examples (including 15k negative examples).

2.2 Estimating the PDM

A point distribution model (PDM) [6] is used for a parametric representation of the non-rigid shape variation in the CLM. The non-rigid warp function can be described as,

$$\mathcal{W}(\mathbf{z}; \mathbf{p}) = \mathbf{z} + \mathbf{V}\mathbf{p} \tag{2}$$

where $\mathbf{z} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T]^T$, $\mathbf{p}$ is a parametric vector describing the non-rigid warp, and $\mathbf{V}$ is the matrix of concatenated eigenvectors. N is the number of patch-experts. Please note that this PDM notation differs slightly from the canonical one because $\mathbf{z}$ is not necessary the mean shape such as defined in [2]. Procrustes analysis [6] is applied to all shape training observations in order remove all similarity. Principal component analysis (PCA) [4] is then employed

to obtain shape eigenvectors $\mathbf{V}$ that preserved 95% of the similarity normalized shape variation in the train set. In this paper, the first 4 eigenvectors of $\mathbf{V}$ are forced to correspond to similarity (i.e., translation, scale and rotation) variation.

3 Constrained Local Model Fitting

Based on the patch experts learned and the point distribution model in Section 2, we can pose non-rigid alignment as the following optimization problem,

$$\arg\min_{\mathbf{p}} \sum_{k} E_k\{Y(\mathbf{x}_k + \mathbf{V}_k\mathbf{p})\} \tag{3}$$

where $E_k()$ is the inverted classifier score function obtained from applying the kth patch expert to the source image patch intensity $Y(\mathbf{x}_k + \Delta\mathbf{x}_k)$. The displacement $\Delta\mathbf{x}_k$ is constrained to be consistent with the PDM defined in Equation 2, where the matrix $\mathbf{V}$ can be decomposed into submatrices $\mathbf{V}_k$ for each kth patch expert, i.e., $\mathbf{V} = [\mathbf{V}_1, \ldots, \mathbf{V}_N]^T$.

One potential problem with the above constrained local model is that the tracking performance is largely dependent on the discriminant performance of the generic patch experts learned in Section 2.1, and there is no guarantee that the alignment results will be consistent between different frames of the same sequence. In order to address this issue, we can extend Equation 3 into the spatio-temporal domain to include the local appearance consistency constraint between neighboring frames. Furthermore, inspired by the approach developed by Baker et al. [3,2], we compute the image error between the input image and the aligned images from previous frames. In particular, we extend Equation 3 as follows

$$\begin{aligned} \arg\min_{\mathbf{p}} \quad & \sum_{k} E_k\{Y(\mathbf{x}_k + \mathbf{V}_k\mathbf{p})\} \\ + \frac{1}{N_{T_0}} & \sum_{t\in T_0} \sum_{k} \lambda_{(t)k} \| Y(\mathbf{x}_k + \mathbf{V}_k\mathbf{p}) - Y_{(t)}(\mathbf{x}_{(t)k}) \|^2 \end{aligned} \tag{4}$$

where $T_0 = [t_0 - \Delta t, t_0]$ is the time interval used to check the local appearance consistence between the current frame Y and the aligned image $Y_{(t)}$ from the previous frame at time t. N_{T_0} [3] is the number of frames included in T_0. $\lambda_{(t)k}$ is the weighting coefficient for the appearance consistency constraint term which is estimated dynamically in Section 5. For clarity, in the rest of this paper we refer to the first term in Equation 4 as the *generic* term and the second one as the *consistency* term.

[3] In our experiments, we typically include 3 previous frames in the appearance consistency constraint term, i.e., $N_{T_0} = 3$.

4 Convex Optimization

In general, it is difficult to solve for $\mathbf{p}$ in Equation 4 as there is no guarantee for the classifier score function $E_k()$ being convex. Previous methods have either used general purpose optimizers (e.g., Nelder-Mead simplex [20]) or attempted to pose the problem as a form of graph optimization [8,14]. Unfortunately, general purpose optimization techniques, such as Nelder-Mead simplex [20], are often computationally expensive and require good initialization. In order to employ graph optimization techniques like loopy belief propagation it has been shown that the warp function $\mathcal{W}(\mathbf{z};\mathbf{p})$ needs to be spatially sparse as described in [14]. In this section, we propose a new approach to jointly optimize $\mathbf{p}$ by convex quadratic fitting.

4.1 Solving the Consistency Term

Since each error function in the consistency term in Equation 4 takes the form of a sum of squared differences (SSD), it can be solved efficiently by the Lucas-Kanade gradient descent algorithm [19,6,2]. For simplicity, we consider the local appearance consistency error function for the kth patch between the current frame Y and the aligned image $Y_{(t)}$ from a previous frame t,

$$\arg\min_{\mathbf{p}} \|Y_{(t)}(\mathbf{x}_{(t)k}) - Y(\mathbf{x}_k + \mathbf{V}_k\mathbf{p})\|^2 \tag{5}$$

where $\mathbf{V}$ is the matrix of concatenated eigenvectors describing the PDM in Equation 2 and $\mathbf{V}_k$ is the submatrix of $\mathbf{V}$ for the kth patch. $\mathbf{p}$ is a parametric vector describing the non-rigid warp.

By performing a first order Taylor series approximation at $Y(\mathbf{x}_k + \mathbf{V}_k\mathbf{p})$, we can rewrite Equation 5 as,

$$\arg\min_{\mathbf{p}} \|D(\mathbf{x}_k) - G^T(\mathbf{x}_k)\mathbf{V}_k\mathbf{p}\|^2 \tag{6}$$

which can be expressed generically in the form of a quadratic,

$$\mathbf{p}^T\mathbf{V}_k^T\mathbf{A}_{(t)k}\mathbf{V}_k\mathbf{p} - 2\mathbf{b}_{(t)k}^T\mathbf{V}_k\mathbf{p} + c_{(t)k} \tag{7}$$

given,

$$\begin{aligned} \mathbf{A}_{(t)k} &= G(\mathbf{x}_k)G^T(\mathbf{x}_k) \\ \mathbf{b}_{(t)k} &= G(\mathbf{x}_k)D(\mathbf{x}_k) \\ c_{(t)k} &= D^T(\mathbf{x}_k)D(\mathbf{x}_k) \end{aligned} \tag{8}$$

where $D(\mathbf{x}_k) = Y_{(t)}(\mathbf{x}_{(t)k}) - Y(\mathbf{x}_k)$ and $G(\mathbf{x}_k)$ is the $2 \times P^2$ local gradient

matrix $\frac{\partial Y(\mathbf{x})}{\partial \mathbf{x}}$ for each set of P^2 intensities centered around $\mathbf{x}_k$.

Therefore, the original consistency term in Equation 4 can be rewritten as

$$\frac{1}{N_{T_0}} \sum_{t \in T_0} \left(\mathbf{p}^T \mathbf{V}^T \mathbf{A}_{(t)} \mathbf{V} \mathbf{p} - 2\mathbf{b}_{(t)}^T \mathbf{V} \mathbf{p} + \mathbf{c}_{(t)} \right) \tag{9}$$

where,

$$\begin{aligned} \mathbf{A}_{(t)} &= \begin{bmatrix} \lambda_{(t)1}\mathbf{A}_{(t)1} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \lambda_{(t)N}\mathbf{A}_{(t)N} \end{bmatrix} \\ \mathbf{b}_{(t)} &= [\lambda_{(t)1}\mathbf{b}_{(t)1}^T, \dots, \lambda_{(t)N}\mathbf{b}_{(t)N}^T]^T \\ \mathbf{c}_{(t)} &= [\lambda_{(t)1}c_{(t)1}, \dots, \lambda_{(t)N}c_{(t)N}]^T \end{aligned}$$

Since each $\mathbf{A}_{(t)k}$ is virtually always guaranteed of being positive definite [4] and the summation of a set of convex functions is still a convex function [5], this implies the quadratic in Equation 9 is convex and has a unique minima given $\lambda_{(t)k} \geq 0$.

4.2 Solving the Generic Term

When assuming $E_k()$ is a SSD classifier it is possible to gain a convex quadratic approximation to the true error responses. A major advantage of these approximations is that it gives a direct method to gain an estimate of the global warp update. In this section we shall elucidate upon how we can generalize this result for any type of objective error function.

Specifically, our approach shall attempt to estimate the parameters $\mathbf{A}_k$, $\mathbf{b}_k$ and c_k, for each patch response surface, through the following optimization

$$\begin{aligned} &\arg \min_{\mathbf{A}_k, \mathbf{b}_k, c_k} \textstyle\sum_{\Delta \mathbf{x}} \| E_k(\Delta \mathbf{x}) \\ &\qquad -\Delta \mathbf{x}^T \mathbf{A}_k \Delta \mathbf{x} + 2\mathbf{b}_k^T \Delta \mathbf{x} - c_k \|^2 \\ &\textit{subject to} \quad \mathbf{A}_k \succ 0 \end{aligned} \tag{10}$$

where $E_k(\Delta \mathbf{x}) = E_k\{Y(\mathbf{x}_k + \Delta \mathbf{x})\}$. We should emphasize that $E_k()$ is now not necessarily a SSD classifier but can be any function that gives a low value for

[4] Actually, $\mathbf{A}_{(t)k}$ is always guaranteed of being positive semidefinite. In the rare occurrence that $\mathbf{A}_{(t)k}$ is positive semidefinite but not positive definite (i.e., singular) we can employ a weighted identity matrix to ensure its rank.

correct alignment. We should note that our proposed approach differs from the standard Lucas-Kanade algorithm in the sense that the actual error response for different translations must be estimated over a local region. In the original Lucas-Kanade approach no such local search responses are required.

After we estimate $\mathbf{A}_k, \mathbf{b}_k$, and c_k in Equation 10 for each patch response surface, the original *generic* term in Equation 4 can be rewritten as

$$\begin{aligned} &\Delta\mathbf{z}^T\mathbf{A}_d\Delta\mathbf{z} - 2\mathbf{b}_k^T\Delta\mathbf{z} + \mathbf{c}_d \\ =&\mathbf{p}\mathbf{V}^T\mathbf{A}_d\mathbf{V}\mathbf{p} - 2\mathbf{b}_d^T\mathbf{V}\mathbf{p} + \mathbf{c}_d \end{aligned} \tag{11}$$

where,

$$\mathbf{A}_d = \begin{bmatrix} \mathbf{A}_1 & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{A}_N \end{bmatrix}$$

$$\mathbf{b}_d = [\mathbf{b}_1^T, \dots, \mathbf{b}_N^T]^T$$

$$\mathbf{c}_d = [c_1, \dots, c_N]^T$$

and $\mathbf{V}$ is the matrix of concatenated eigenvectors describing the PDM in Equation 2. We shall refer to this method of fitting a CLM as *convex quadratic fitting* (CQF). The keypoint of enforcing the convexity of each local patch response is to find a convex local function, which is essential to achieve a fast convergence for the global optimization. The detailed computational complexity analysis can be found in [24].

Exhaustive Local Search: Rather than solving Equation 10, for computational efficiency it often convenient to assume that $\mathbf{A}_k = \sigma\mathbf{I}$ where $\sigma \to 0$ and that,

$$\mathbf{b}_k = \arg\min_{\Delta\mathbf{x}} E_k(\Delta\mathbf{x}) \tag{12}$$

which results in finding the minimum point for each response map. We refer to this approach as exhaustive local search (ELS). ELS is equivalent to how an ASM performs its fitting procedure [7] and is a good baseline from which to compare other fitting strategies.

Quadratic Program Curve Fitting: The optimization in Equation 10 is in general costly if solved directly [5]. One way to reduce the complexity of Equation 10 is to enforce $\mathbf{A}_k$ to be a diagonal matrix with non-negative diagonal elements. More specifically, for 2D image alignment $\mathbf{A}_k = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix}$

where $a_{11}, a_{22} > 0$. As a result, Equation 10 can be simplified as

$$\begin{aligned} &\arg\min_{a_{11},a_{22},b_1,b_2,c} \textstyle\sum_{x,y} \|E_k(x,y) \\ &-a_{11}x^2 - a_{22}y^2 + 2b_1x + 2b_2y - c\|^2 \\ &\textit{subject to} \quad a_{11} > 0, a_{22} > 0 \end{aligned} \tag{13}$$

which can be solved efficiently through quadratic programming [5].

Robust Error Function: When the local search responses from our patch experts have outliers, it might be difficult to have accurate surface fitting. To address this issue, robust error functions have been used in many registration approaches [2,21] to improve robustness for non-rigid image alignment. Although there are many different choices [21], a sigmoid function is selected similar to the weighting function in Equation 15. In particular, we define the robust error function in the following form,

$$\varrho(\mathcal{E}(\mathbf{x}); \sigma) = \frac{1}{1 + e^{-\|\mathcal{E}(\mathbf{x})\|^2 + \sigma}}$$

where σ is a scale parameter which can be estimated from $\mathcal{E}(\mathbf{x})$. Essentially, this function assigns lower weights to the response values whose fitting error is larger than the scale parameter σ, since they are more likely to be the outliers. As a result, the original curve fitting problem in Equation 10 can be rewritten as

$$\begin{aligned} &\arg\min_{\mathbf{A}_k,\mathbf{b}_k,c_k} \textstyle\sum_{\Delta\mathbf{x}} \varrho(\mathcal{E}(\Delta\mathbf{x}); \sigma) \\ &\textit{subject to} \quad \mathbf{A}_k \succ 0 \end{aligned} \tag{14}$$

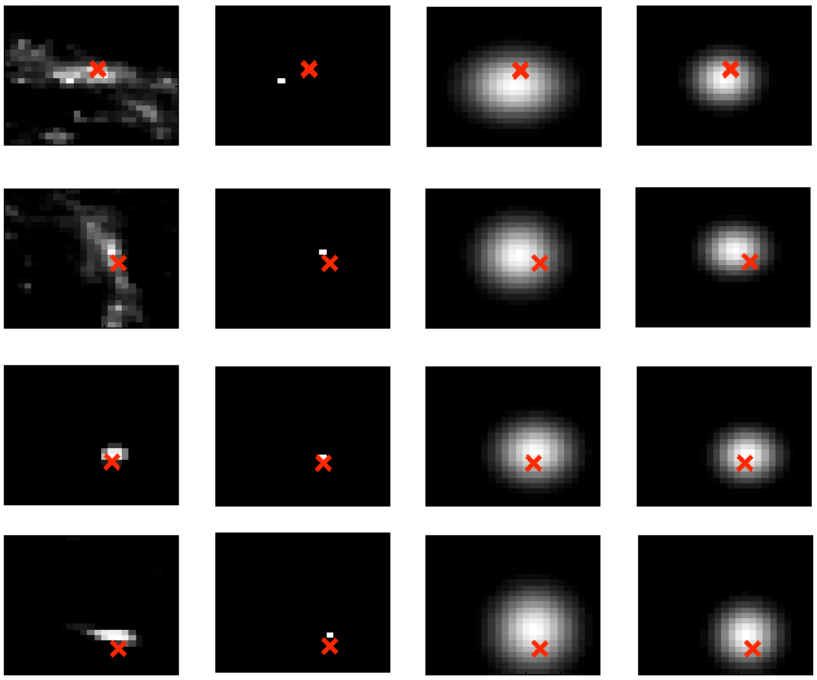
where

$$\mathcal{E}(\Delta\mathbf{x}) = E(\Delta\mathbf{x}) - \Delta\mathbf{x}^T\mathbf{A}_k\Delta\mathbf{x} + 2\mathbf{b}_k^T\Delta\mathbf{x} - c_k.$$

We shall refer to this method of fitting a CLM as *robust convex quadratic fitting* (RCQF) [24].

Example Fits: Examples of local response surface fitting can be found in Figure 2, which illustrates the convex parametric representation of the non-parametric responses of local patch experts. The red cross shows the ground truth location in the search window. The closer the peaks of the local responses are to the red cross indicates the better the performance of the method. We can see that in most cases ELS, CQF, and RCQF methods can all achieve good performance. However, our proposed CQF and RCQF methods in (c) and (d) respectively are less sensitive to local minima than the ELS method in (b). We should note that although the learned patch responses look smooth, they are not generated by a mere smoothing step. Instead, they are continuous convex surfaces achieved by the constrained curve fitting proposed in this paper. The key point of enforcing the convexity of each local patch response is to find a

convex local function, which is essential to achieve a fast convergence for the global optimization.



(a) (b) (c) (d)

Fig. 2. Examples of fitting local search responses: (a) is the local search responses in Figure 1(d) using patch experts trained by a linear support vector machine (SVM). (b-d) show the surface fitting results. More specifically, (b) picks the local displacement with the minimum response value in the search window, while (c) and (d) fit the local search response surface by a quadratic kernel in Equation 13 and a quadratic kernel with a robust error function in Equation 14, respectively. The brighter intensity means the smaller matching error between the template and the source image patch. In each search window, the red cross illustrates the ground truth location. As we can see, in most cases, the above three methods can all achieve good performance, while the proposed convex quadratic fitting (CQF) (c) and the robust convex quadratic fitting (RCQF) (d) methods are less sensitive to local minima than the exhaustive local search (ELS) method (b).

5 Estimating Weights

The choice of each weighting coefficient $\lambda_{(t)k}$ plays an important role in obtaining the optimal solution of Equation 4. A small value might not be able to impose enough smoothness constraints on the tracking results while a large value might cause other issues such as drifting. To address this issue, we can estimate the weight values $\lambda_{(t)k}$ dynamically based on how likely the aligned patches extracted from the previous frames are good templates. Although we can measure the quality of match by introducing certain prior model such as

in [22], a simple approach is to update the weights based on the output $\hat{f}$ of the support vector machine from Equation 1.

More specifically, an approximate probabilistic output can be obtained by fitting a logistic regression function [4] to the output $\hat{f}$ of Equation 1 and the labels $y = \{\text{not aligned } (-1), \text{aligned } (+1)\}$

$$\hat{P}(y = 1|\hat{f}) = \frac{1}{1 + e^{a\hat{f}+b}} \tag{15}$$

where a and b are learned through a cross-validation process. Then we define $\lambda_{(t)k}$ using the approximate probabilistic output $\hat{P}(y = 1|\hat{f}_{(t)k})$ as follows

$$\begin{aligned} \lambda_{(t)k} &= \eta \left(1 - \hat{P}(y = 1|\hat{f}_{(t)k})\right) \\ &= \frac{\eta e^{a\hat{f}_{(t)k}+b}}{1 + e^{a\hat{f}_{(t)k}+b}} \end{aligned} \tag{16}$$

where

$$\hat{f}_{(t)k} = Y_{(t)}(\mathbf{x}_{(t)k})^T \sum_{i=1}^{N_S} \gamma_i \alpha_i T_i(\mathbf{x}_k)$$

where $Y_{(t)}$ is the aligned image of the frame t, T_i is the ith learned support vector, γ_i is the corresponding support label, α_i is the corresponding support weight and N_S is the number of support vectors. The intuition behind Equation 16 is that the consistency term only comes to help when the associated patch experts can not locate the feature points correctly, i.e., the SVM score $\hat{f}_{(t)k}$ is low.

As discussed in Section 2.1, Equation 16 can be computed efficiently because of the advantageous property of a linear SVM, which allows for $\sum_{i=1}^{N_S} \gamma_i \alpha_i T_i(\mathbf{x})$ to be pre-computed rather than evaluated at every frame. a and b are the same as in Equation 15 and η is learned through a cross-validation process. As shown in Figure 4, the choice of η does not have a significant affect on the tracking performance of our proposed method. In our experiments, we typically set η a small value 0.1.

6 Our Algorithm

A major advantage of the convex quadratic fitting (CQF) method proposed in Section 4.2 is that it makes both the generic term and the consistency term in Equation 4 share the same quadratic form. As a result, we can simplify the original optimization problem in Equation 4 and solve jointly for the global non-rigid shape of the object in an efficient manner. More specifically, based

on Equation 9 and 11 we can rewrite Equation 4 as follows,

$$\begin{aligned}\arg\min_{\mathbf{p}} \quad & \mathbf{p}^T\mathbf{V}^T\mathbf{A}_d\mathbf{V}\mathbf{p} - 2\mathbf{b}_d^T\mathbf{V}\mathbf{p} + \mathbf{c}_d \\ + \frac{1}{N_{T_0}} & \sum_{t\in T_0}(\mathbf{p}^T\mathbf{V}^T\mathbf{A}_{(t)}\mathbf{V}\mathbf{p} - 2\mathbf{b}_{(t)}^T\mathbf{V}\mathbf{p} + \mathbf{c}_{(t)}) \\ = \arg\min_{\mathbf{p}} \quad & \mathbf{p}^T\mathbf{V}^T\mathbf{A}\mathbf{V}\mathbf{p} - 2\mathbf{b}^T\mathbf{V}\mathbf{p} + \mathbf{c} \end{aligned} \tag{17}$$

where,

$$\begin{aligned}\mathbf{A} &= \mathbf{A}_d + \tfrac{1}{N_{T_0}}\textstyle\sum_{t\in T_0}\mathbf{A}_{(t)} \\ \mathbf{b} &= \mathbf{b}_d + \tfrac{1}{N_{T_0}}\textstyle\sum_{t\in T_0}\mathbf{b}_{(t)} \\ \mathbf{c} &= \mathbf{c}_d + \tfrac{1}{N_{T_0}}\textstyle\sum_{t\in T_0}\mathbf{c}_{(t)}\end{aligned}$$

where $\mathbf{V}$ is the matrix of concatenated eigenvectors describing the PDM defined as in Equation 2, $\mathbf{p}$ is a parametric vector describing the non-rigid warp, N is the number of patch-experts, and $(\mathbf{A}_d, \mathbf{b}_d,\ \mathbf{c}_d)$ and $(\mathbf{A}_{(t)}, \mathbf{b}_{(t)},\ \mathbf{c}_{(t)})$ are defined in Equation 11 and 9 respectively.

Furthermore, as discussed in Section 4.1 and 4.2 $\mathbf{A}_d$ and $\mathbf{A}_{(t)}$ are both positive definite. Since the summation of a set of convex functions is still a convex function [5], given $\lambda_{(t)k} \geq 0$ it is possible to solve not only for the local translation updates but the entire warp update $\mathbf{p}$ explicitly,

$$\mathbf{p} = \left(\mathbf{V}^T\mathbf{A}\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{b} \tag{18}$$

where $\mathbf{V}$ is the matrix of concatenated eigenvectors describing the PDM defined in Equation 2.

When the robust error functions are applied to the CLM fitting as in Equation 14, by performing a first-order Taylor expansion of $\varrho(\mathcal{E}(\Delta\mathbf{x});\sigma)$, we can derive the global update $\Delta\mathbf{p}$ explicitly in a similar form to Equation 18 where

$$\begin{aligned}\mathbf{A} &= \mathbf{B}\mathbf{A}_d + \tfrac{1}{N_{T_0}}\textstyle\sum_{t\in T_0}\mathbf{A}_{(t)} \\ \mathbf{b} &= \mathbf{B}\mathbf{b}_d + \tfrac{1}{N_{T_0}}\textstyle\sum_{t\in T_0}\mathbf{b}_{(t)}\end{aligned}$$

$\mathbf{B}$ is a $2N \times 2N$ diagonal matrix with

$$\begin{aligned}\mathbf{B}_{(i,i)} &= \frac{\partial\varrho(\mathcal{E}(x_k, y_k);\sigma_k)}{\partial x_k} \\ \mathbf{B}_{(i+1,i+1)} &= \frac{\partial\varrho(\mathcal{E}(x_k, y_k);\sigma_k)}{\partial y_k}\end{aligned}$$

where $i = 2k$ and $k = 1 \ldots N$.

Input:- learned patch experts, source image (Y),
aligned images from the previous frames ($Y_{(t)}$),
Jacobian matrix ($\mathbf{V}$),
initial warp guess ($\mathbf{p}$),
index to the template ($\mathbf{z}$), threshold (ϵ)
Output:- final warp ($\mathbf{p}$)

(1) Warp the source image Y with the current similarity transform from $\mathbf{p}$.

(2) Compute the local responses E based on the learned patch experts and the source image Y.

(3) Estimate the convex quadratic curve fitting parameters $\mathbf{A}_k$, $\mathbf{b}_k$ and c_k from Equation 13 for each patch.

(4) Compute the weights $\lambda_{(t)k}$ using Equation 16.

(5) Estimate the warp update $\Delta\mathbf{p}$ using Equation 18.

(6) Update the warp $\mathbf{z}' = \mathcal{W}(\mathbf{z};\mathbf{p})$ using $\mathcal{W}(\mathbf{z};\mathbf{p}) \leftarrow \mathcal{W}(\mathbf{z};\mathbf{p}) \circ \mathcal{W}(\mathbf{z};\Delta\mathbf{p})$.

(7) Repeat steps 1-6 until $||\Delta\mathbf{p}|| <= \epsilon$ or max iterations reached.

Algorithm 1. The outline of our spatio-temporal convex quadratic fitting (ST-CQF) method.

Since we are only using an approximation to the true SSD error surface it is necessary within the Lucas-Kanade algorithm to iterate this operation and constantly update the warp estimate $\mathbf{p}$ until convergence. For clarity, we list the outline of our spatio-temporal convex quadratic fitting (ST-CQF) method in Algorithm 6.

7 Experiments

We conducted our experiments on a clinical archive, which contains video clips of clinical patients with shoulder injuries. These clips have a large amount of head motion and facial expressions. All the images had 66 fiducial points annotated as the ground truth data. To make this task even more challenging, we trained all models, including the PDM and the patch experts, separately on the MultiPIE face database [12] which does not include any subjects from the clinical archive.

7.1 Evaluation

In all our experiments the similarity normalized base template had an inter-ocular distance of 50 pixels. For a fair comparison, we took into account differing face scales between testing images. This is done by first removing the similarity transform between the estimated shape and the base template shape and then computing the RMS-PE between the 66 points. To compare the performance of different algorithms we employed an *alignment convergence curve* (ACC) [8]. These curves have a threshold distance in RMS-PE on the x-axis and the percentage of trials that achieved convergence (i.e., final alignment RMS-PE below the threshold) on the y-axis. A perfect alignment algorithm would receive an ACC that has 100% convergence for all threshold values.

7.2 Comparison Results

In this section we evaluate the performance of our proposed algorithm to track non-rigid facial motion in video sequences. To evaluate the performance we conducted comparison experiments on a subset of a clinical archive which included 22 video clips of 10 clinical patients with significant head motion and facial expressions. There are $200 - 400$ frames in each video sequence. We trained all models, including the PDM and the patch experts, separately on the MultiPIE face database [12]. Since no subjects are shared between the training and testing databases, the appearance and shape variances are very different between them which makes the face alignment/tracking task a very challenging problem. For completeness, we also included the *simultaneous* AAM method which is considered one of the leading algorithms for holistic non-rigid alignment [2]. In our results we shall refer to this algorithm simply as the AAM method. Figure 3 shows the results of our comparison.

As discussed in Section 1, the CLM methods have several advantages over the holistic AAM method in terms of accuracy and robustness to appearance variation. The results in Figure 3 on the clinical archive further support these claims. We can see in Figure 3 that the CLM algorithms all outperformed the AAM method. Furthermore, the *spatio-temporal convex quadratic fitting* (ST-CQF) method proposed in Section 6 received better performance than both the *robust convex quadratic fitting* (RCQF) and *convex quadratic fitting* (CQF) methods by integrating the local appearance constraint. One hypothesis is that the patch experts trained in one data set does not perform as well in a new data set. By enforcing the local appearance consistency constraint, the joint optimization can reduce the local appearance ambiguity and improve the robustness and accuracy of the non-rigid alignment.

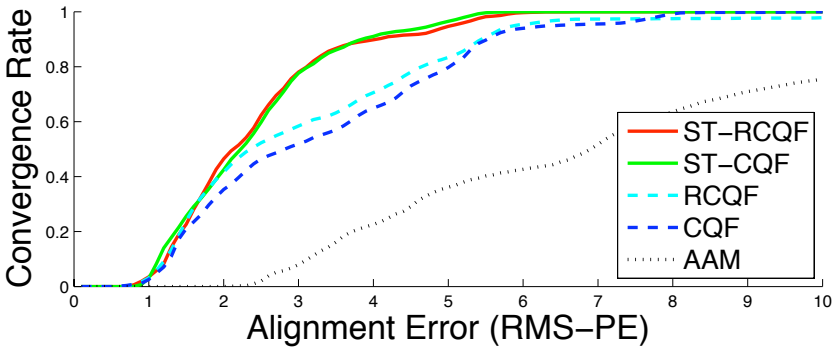


Fig. 3. A comparison of tracking results for 22 video clips of 10 pain patients with significant head motion and facial expression. Each video has $200 - 400$ frames. We trained all models, including the PDM and the patch experts, separately on the MultiPIE face database [12]. Three methods were included in the comparison: (i) spatio-temporal convex quadratic fitting (ST-CQF), (ii) convex quadratic fitting (CQF) and (iii) active appearance model (AAM). ST-CQF and CQF with robust error functions, i.e., ST-RCQF and RCQF, were also included in the comparison experiments. The weighting scale factor η was 0.1 in both ST-CQF and ST-RCQF. Conforming to Section 1, the CLM methods all outperformed the holistic AAM method in terms of higher alignment accuracy and convergence rates. Furthermore, the proposed ST-CQF method had better alignment performance than both the RCQF and CQF methods.

An interesting observation in Figure 3 is that there is not much difference between the performance of ST-CQF and ST-RCQF. One potential explanation is that the temporal texture consistency constraints greatly remove the outliers occurred to the local patch-expert matching, which improves the robustness of the object alignment in a similar way as the robust error functions. Therefore the proposed ST-CQF method can achieve accurate and robust object tracking performance without using the computationally expensive robust error functions. Examples of alignment result on different subjects are also shown in Figure 5 and 6 to illustrate the performance of the three different methods compared in Figure 3(a).

Furthermore, as described in Section 5, the weights for the consistency term in the overall objective error function 4 is computed based on the parameter η in Equation 16. To analyze how sensitive the performance of our proposed tracking method is to the value of η, we also conducted comparison experiments with a wide range of η values. The results are reported in Figure 4. The proposed spatio-temporal convex quadratic fitting (ST-CQF) method with different η values all had much better performance than the convex quadratic fitting (CQF) method without the temporal appearance consistency constraint (i.e., $\eta = 0$). Furthermore, the choice of different weights η does not have a significant affect the tracking performance of our proposed method.

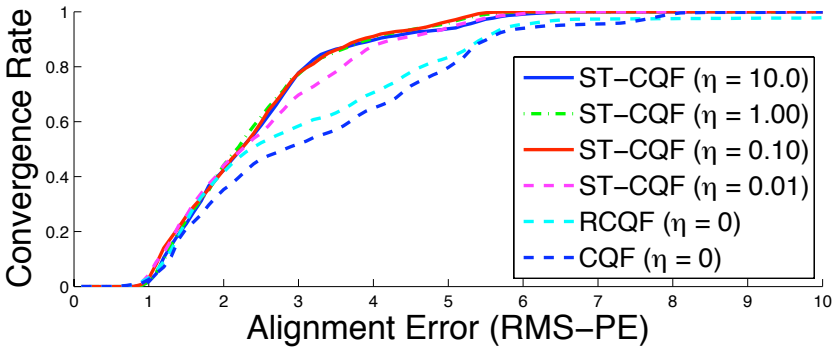
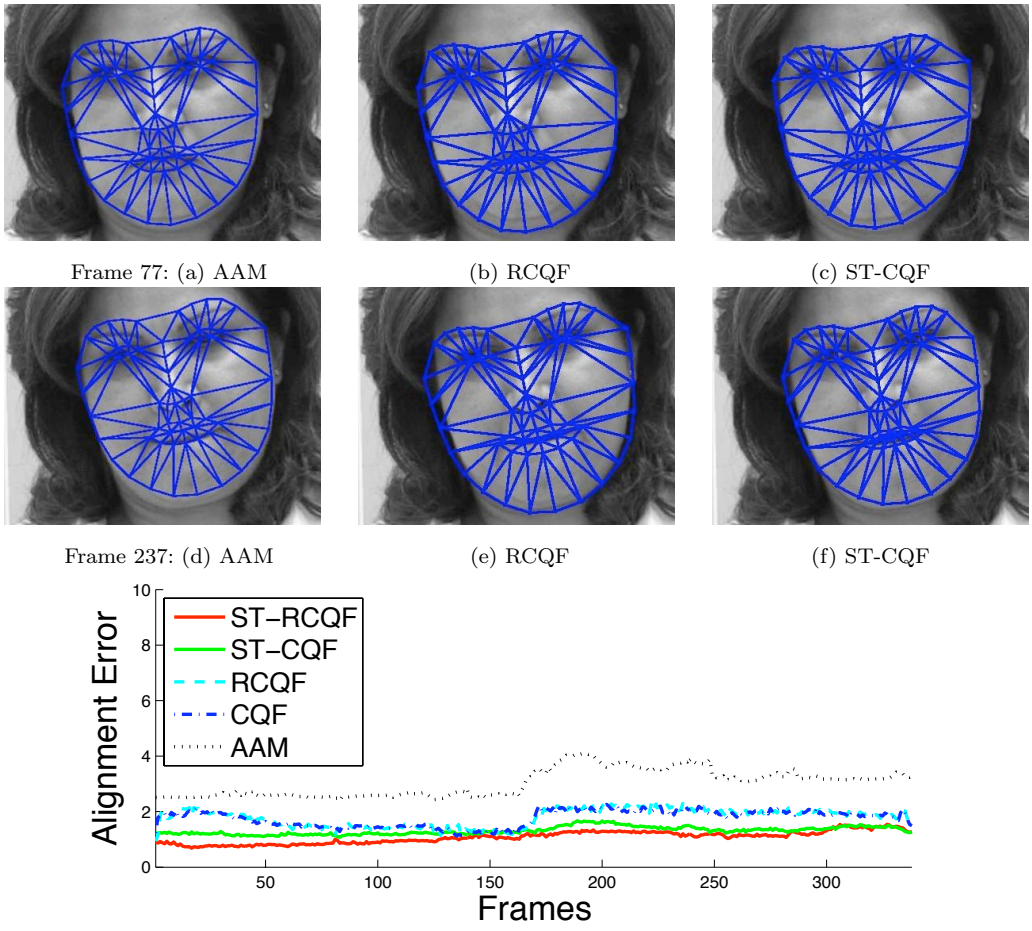


Fig. 4. A comparison of tracking results with different weights η for the consistency term. The same training and testing dataset were used as described in the caption of Figure 3. The proposed spatio-temporal convex quadratic fitting (ST-CQF) method had much better performance than both the robust convex quadratic fitting (RCQF) and the convex quadratic fitting (CQF) methods. Furthermore, the choice of different weights η does not have a significant affect to the tracking performance of our proposed method.

8 Conclusion and Future Work

In this paper, we proposed a new discriminative approach to tracking non-rigid object motion, such as facial expressions, in an efficient and unsupervised manner. By extending the canonical constrained local models (CLM) framework [8] into the spatio-temporal domain, the proposed approach can reduce ambiguity and increase accuracy. Furthermore, we formulated the optimization problem into a convex quadratic curve fitting framework whose generic term and consistency term share the same quadratic form. This convex quadratic framework was motivated by the effectiveness of the canonical Lucas-Kanade algorithm when dealing with a similar optimization problem. By enforcing this convexity it was possible, through an iterative method, to solve jointly for the global non-rigid shape of the object.

We evaluated the performance of our proposed method using the videos from a clinical archive which contains video clips of pain patients. The experimental results demonstrated that our spatio-temporal convex quadratic (ST-CQF) CLM has better alignment performance than other evaluated CLMs without the local appearance consistency constraint and leading existing holistic methods for alignment/tracking (i.e., AAMs). In future work, we shall investigate other discriminant classifiers such as boosting schemes [4,18] or relevance vector machine (RVMs) [4] to further improve the performance of our patch experts. We would also like to explore alternate geometric constraints to handle large deformations and occlusion.



Frame 77: (a) AAM (b) RCQF (c) ST-CQF

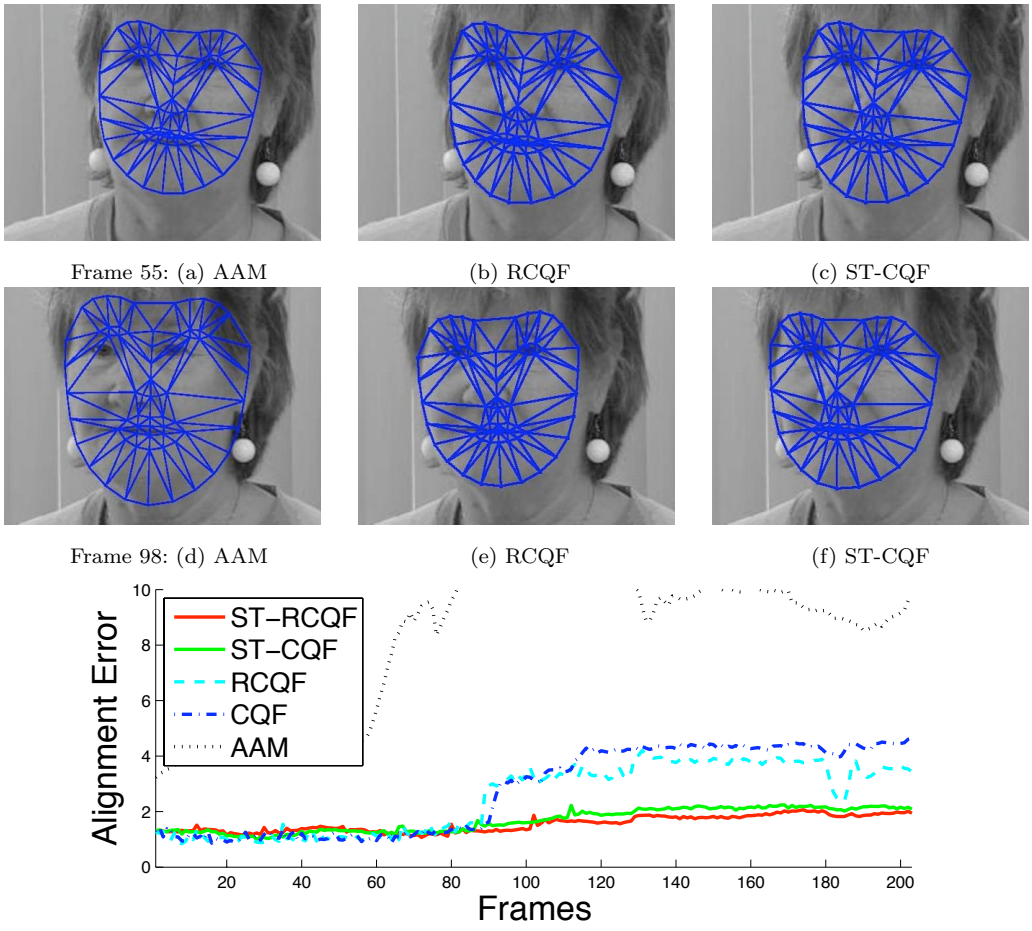
Frame 237: (d) AAM (e) RCQF (f) ST-CQF

(g) Tracking Error Comparison

Fig. 5. Examples of tracking performance on an unseen facial expression sequence. Since the MultiPIE face database [12] does not include the lip tightening expression, the appearance variation around the lips was not included in the training dataset. There are 338 frames in the sequence and the first and second row shows the tracking results of the 77th and 237 frame, respectively. The first column (a,d) shows the resulting alignment from the holistic active appearance model (AAM), the second column (b,e) from the robust convex quadratic fitting (RCQF), and the third column (c,f) from our spatio-temporal convex quadratic fitting (ST-CQF) method. The plot in the third row shows the comparison of tracking error (RMS-PE) on each frame of the whole sequence between the 5 methods as described in Figure 3, i.e., AAM, CQF, RCQF, ST-CQF and ST-RCQF. The weighting scale factor η was set as 0.1 in both ST-CQF and ST-RCQF. Since this facial expression was not included in the training database, the learned appearance model could not find good matching around the lips even with the help of robust error functions. However, our proposed ST-CQF and ST-RCQF methods can achieve a good alignment performance by enforcing the local appearance consistency in the temporal domain.

9 Acknowledgements

The authors would like to thank Mark Cox for the helpful discussions. This work was partially supported by NIH Grant R01 MH051435 and CIHR Grant



(g) Tracking Error Comparison

Fig. 6. Comparison experiments on drifting. There are 202 frames in the sequence and the first and second row shows the tracking results of the 55th and 98th frame, respectively. The plot in the bottom row shows that both the RCQF and CQF methods started to drift around the 90th frame while our proposed ST-CQF method can maintain a consistent tracking performance with a high accuracy. The first column (a,d) shows the resulting alignment from the holistic active appearance model (AAM), the second column (b,e) from the robust convex quadratic fitting (RCQF), and the third column (c,f) from our spatio-temporal convex quadratic fitting (ST-CQF) method. The plot in the third row includes the comparison of tracking error (RMS-PE) through the whole sequence between the 5 methods as described in Figure 3, i.e., AAM, CQF, RCQF, ST-CQF and ST-RCQF. The weighting scale factor η was 0.1 in both ST-CQF and ST-RCQF. Our proposed ST-CQF and ST-RCQF methods had much more accurate and temporally smoother tracking results than both CQF and RCQF methods.

145703.

References

[1] S. Avidan. Support vector tracking. *PAMI*, 26(8):1064–1072, August 2004.

[2] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 1: The quantity approximated, the warp update rule, and the gradient descent approximation. *IJCV*, 2004.

[3] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *PAMI*, 26(10):1380–1384, October 2004.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] S Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV*, volume 2, pages 484–498, 1998.

[7] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models - their training and applications. , 61(2), january 1995. *Computer Vision and Image Understanding*, 61(2), 1995.

[8] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, pages 929–938, 2006.

[9] N.D.H. Dowson and R. Bowden. N-tier simultaneous modelling and tracking for arbitrary warps. In *BMVC*, page II:569, 2006.

[10] P.F. Felzenszwalb and D.P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, January 2005.

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

[12] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. The CMU Multiple pose, illumination and expression (MultiPIE) database. Technical Report CMU-RI-TR-07-08, Robotics Institute, Carnegie Mellon University, 2007.

[13] L. Gu and T. Kanade. 3D Alignment of face in a single image. In *CVPR*, volume 1, pages 1305–1312, 2006.

[14] L. Gu, E.P. Xing, and T. Kanade. Learning gmrf structures for spatial priors. In *CVPR*, pages 1–6, 2007.

[15] I. Kokkinos and A.L. Yuille. Unsupervised learning of object deformation models. In *ICCV07*, pages 1–8, 2007.

[16] E.G. Learned Miller. Data driven image models through continuous joint alignment. *PAMI*, 28(2):236–250, 2006.

[17] L. Liang, F. Wen, Y.Q. Xu, X. Tang, and H.Y. Shum. Accurate face alignment using shape constrained Markov network. In *CVPR*, pages I: 1313–1319, 2006.

[18] X.M. Liu. Generic face alignment using boosted appearance model. In *CVPR*, pages 1–8, 2007.

[19] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[20] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[21] Barry-John Theobald, Iain Matthews, and Simon Baker. Evaluating error functions for robust active appearance models. In *International Conference on Automatic Face and Gesture Recognition*, pages 149–154, 2006.

[22] K.N. Walker, T.F. Cootes, and C.J. Taylor. Automatically building appearance models from image sequences using salient features. *IVC*, 20(5-6):435–440, 2002.

[23] Y. Wang, S. Lucey, and J. Cohn. Non-rigid object alignment with a mismatch template based on exhaustive local search. In *IEEE Workshop on Non-rigid Registration and Tracking through Learning*, 2007.

[24] Y. Wang, S. Lucey, and J. Cohn. Enforcing convexity for improved alignment with constrained local models. In *CVPR*, 2008.

[25] O. Williams, A. Blake, and R. Cipolla. Sparse Bayesian learning for efficient visual tracking. *PAMI*, 27(8):1292–1304, August 2005.

[26] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+3d active appearance models. In *CVPR*, pages II: 535–542, 2004.

[27] Y. Zhou, L. Gu, and H. Zhang. Bayesian tangent shape model: Estimating shape and pose parameters via Bayesian inference. In *CVPR*, volume 1, pages 109–116, 2003.