Non-rigid Face Tracking with Local Appearance Consistency Constraint

Yang Wang, Simon Lucey, Jeffrey F. Cohn, Jason Saragih
The Robotics Institute, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh PA 15213, USA
{wangy,slucey,jeffcohn}@cs.cmu.edu,jsaragih@andrew.cmu.edu

Abstract

In this paper we present a new discriminative approach to achieve consistent and efficient tracking of non-rigid object motion, such as facial expressions. By utilizing both spatial and temporal appearance coherence at the patch level, the proposed approach can reduce ambiguity and increase accuracy. Recent research demonstrates that feature based approaches, such as constrained local models (CLMs), can achieve good performance in non-rigid object alignment/tracking using local region descriptors and a non-rigid shape prior. However, the matching performance of the learned generic patch experts 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 CLM method into the spatio-temporal domain by enforcing the appearance consistency constraint of each local patch between neighboring frames. More importantly, we show that the global warp update can be optimized jointly in an efficient manner using convex quadratic fitting. *Finally, we demonstrate that our approach receives improved performance for the task of non-rigid facial motion tracking on the videos of clinical patients.*

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 [1,2,6–9,11,15,16,23–25]. This problem is particularly difficult when tracking subjects with 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 [7, 11, 12, 15, 16, 20, 21]. Compared to the holistic representations, such as active appearance models (AAMs), working on the patch level offers us more flexibilities on the local regions, which leads to improved accuracy on object alignment and robustness to unseen appearance variation [7,11,21]. In particular, the constrained local model (CLM) framework proposed recently by Cristinacce and Cootes [7] has demonstrated good performance in non-rigid object alignment/tracking, in comparison to leading holistic approaches (e.g., AAMs). Instead of using holistic representations, a CLM is able to register a non-rigid object through the application of an ensemble of patch/region experts to local search regions within the source image. Given an appropriate non-rigid shape prior for the object, the response surfaces from these local regions are then employed within a joint optimization process to estimate the global non-rigid shape of the object. However, the matching performance of the learned generic 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 object alignment might not be consistent from frame to frame and the motion field is not temporally smooth.

Since there is texture coherence between different images, another direction is to explore this information to align them automatically. Inspired by recent work for aligning a set of images in an unsupervised manner [3, 13, 14, 20] we propose a new discriminative 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 original CLM method 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

978-1-4244-2154-1/08/$25.00

frames. (Section 3)

- Instead of using computationally expensive generic optimizers such as the Nelder-Mead simplex [7] method, we propose a *convex quadratic fitting* 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 5)

- Finally, we demonstrate improved non-rigid alignment 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 6)

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[1] 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 a variety of methods such as boosting schemes [4, 16] (e.g., AdaBoost, GentleBoost, etc.) or relevance vector machine (RVMs) [4] to mention just a few. A linear SVM was chosen in our work over other classifiers due to its computational advantages in that,

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

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. 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 rather than evaluated at every $\Delta\mathbf{x}$. The support images T_i are obtained from an offline 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.

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.

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,

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

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

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

as the following optimization problem,

$$\arg\min_{\mathbf{p}} \sum_k E_k\{Y(\mathbf{x}_k + \mathbf{V}_k\mathbf{p})\} \quad (3)$$

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*. [2, 3], we compute the image error between the input image and the aligned images from previous frames. In particular, we extend Equation 3 as follows

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

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. ${N_{T_0}}^2$ 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 4.3. 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.

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 [18]) or attempted to pose the problem as a form of graph optimization [7, 12]. Unfortunately, general purpose optimization techniques, such as Nelder-Mead simplex [18], 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 [12]. In this section, we propose a new approach to jointly optimize $\mathbf{p}$ by convex quadratic fitting.

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

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 [2, 6, 17]. 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,

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

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,

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

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

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

given,

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

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

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

where,

$$\mathbf{A}_{(t)} = \begin{bmatrix} \lambda_{(t)1}\mathbf{A}_{(t)1} & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \lambda_{(t)N}\mathbf{A}_{(t)N} \end{bmatrix}$$

$$\begin{aligned} \mathbf{b}_{(t)} &= [\lambda_{(t)1}\mathbf{b}_{(t)1}^T, \ldots, \lambda_{(t)N}\mathbf{b}_{(t)N}^T]^T \\ \mathbf{c}_{(t)} &= [\lambda_{(t)1}c_{(t)1}, \ldots, \lambda_{(t)N}c_{(t)N}]^T \end{aligned}$$

Since each $\mathbf{A}_{(t)k}$ is virtually always guaranteed of being positive definite[3] 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

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

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 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

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

where,

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

$$\begin{array}{rcl} \mathbf{b}_d & = & [\mathbf{b}_1^T, \dots, \mathbf{b}_N^T]^T \\ \mathbf{c}_d & = & [c_1, \dots, c_N]^T \end{array}$$

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 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. The detailed computational complexity analysis can be found in [22].

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

$$\begin{array}{rl} \arg\min_{a_{11},a_{22},b_1,b_2,c} & \sum_{x,y} \|E_k(x,y) \\ -a_{11}x^2 - a_{22}y^2 + & 2b_1x + 2b_2y - c\|^2 \\ subject & to \quad a_{11} > 0, a_{22} > 0 \end{array} \tag{12}$$

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, 19] to improve robustness for non-rigid image alignment. Although there are many different choices [19], a sigmoid function is selected similar to the weighting function in Equation 14. 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

$$\begin{array}{rl} \arg\min_{\mathbf{A}_k,\mathbf{b}_k,c_k} & \sum_{\Delta\mathbf{x}} \varrho(\mathcal{E}(\Delta\mathbf{x}); \sigma) \\ subject & to \quad \mathbf{A}_k \succ 0 \end{array} \tag{13}$$

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) [22].

4.3. 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

[3] 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.

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 [20], 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 $y = \{$not aligned (-1), aligned $(+1)\}$

$$\hat{P}(y = 1|\hat{f}) = \frac{1}{1 + e^{a\hat{f}+b}} \tag{14}$$

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

$$\begin{aligned} \lambda_{(t)k} &= \eta\left(1 - \hat{P}(y = 1|\hat{f}_{(t)k})\right) \\ &= \frac{\eta e^{a\hat{f}_{(t)k}+b}}{1 + e^{a\hat{f}_{(t)k}+b}} \end{aligned} \tag{15}$$

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 15 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 15 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 precomputed rather than evaluated at every frame. a and b are the same as in Equation 14 and η is learned through a cross-validation process. As shown in Figure 2, 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.

5. 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,

$$\begin{aligned} &\arg\min_{\mathbf{p}} \quad \mathbf{p}^T\mathbf{V}^T\mathbf{A}_d\mathbf{V}\mathbf{p} - 2\mathbf{b}_d^T\mathbf{V}\mathbf{p} + \mathbf{c}_d \\ &+\frac{1}{N_{T_0}}\sum_{t\in T_0}(\mathbf{p}^T\mathbf{V}^T\mathbf{A}_{(t)}\mathbf{V}\mathbf{p} - 2\mathbf{b}_{(t)}^T\mathbf{V}\mathbf{p} + \mathbf{c}_{(t)}) \\ &= \arg\min_{\mathbf{p}} \ \mathbf{p}^T\mathbf{V}^T\mathbf{A}\mathbf{V}\mathbf{p} - 2\mathbf{b}^T\mathbf{V}\mathbf{p} + \mathbf{c} \end{aligned} \tag{16}$$

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,

$$\mathbf{p} = \left(\mathbf{V}^T\mathbf{A}\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{b} \tag{17}$$

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 12 for each patch.
4. Compute the weights $\lambda_{(t)k}$ using Equation 15.
5. Estimate the warp update $\Delta\mathbf{p}$ using Equation 17.
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 1.

6. 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 [10] which does not include any subjects from the clinical archive.

6.1. 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) [7]. 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.

6.2. 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 [10]. 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 1 shows the results of our comparison.

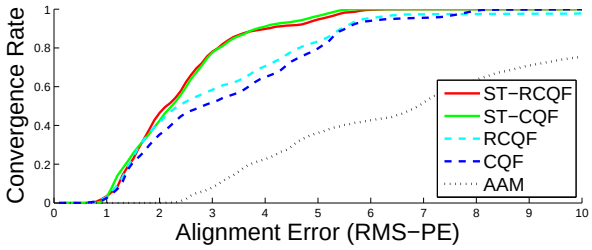


Figure 1. A comparison of tracking results for 22 video clips of 10 pain patients with significant head motion and facial expression. Each video has $200-400$ frames. We trained all models, including the PDM and the patch experts, separately on the MultiPIE face database [10]. Three methods were included 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.

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 1 on the clinical archive further support these claims. We can see in Figure 1 that the CLM algorithms all outperformed the AAM method. Furthermore, the *spatio-temporal convex quadratic fitting* (ST-CQF) method proposed in Section 5 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.

An interesting observation in Figure 1 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 3 and 4 to illustrate the performance of the three different methods compared in Figure 1(a).

Furthermore, as described in Section 4.3, the weights for the consistency term in the overall objective error function 4 is computed based on the parameter η in Equation 15. 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 2. 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.

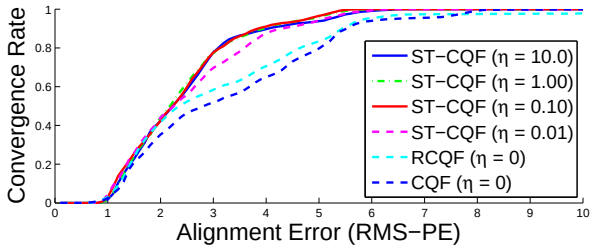


Figure 2. A comparison of tracking results with different weights η for the consistency term. The same training and testing dataset were used as described in the caption of Figure 1. The proposed spatio-temporal convex quadratic fitting (ST-CQF) method had much better performance than both the robust 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.

7. 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 [7] 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, 16] 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.

8. Acknowledgements

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

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

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

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

[10] 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.

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

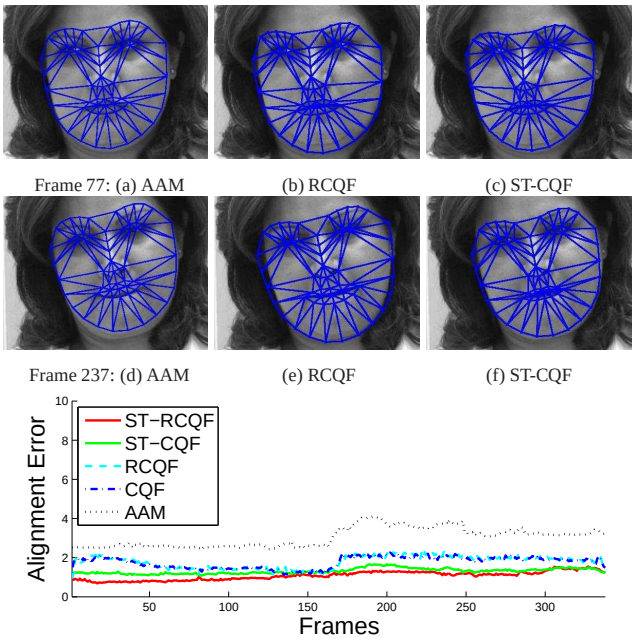
[12] L. Gu, E. Xing, and T. Kanade. Learning gmrf structures for spatial priors. In *CVPR*, pages 1–6, 2007.

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

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

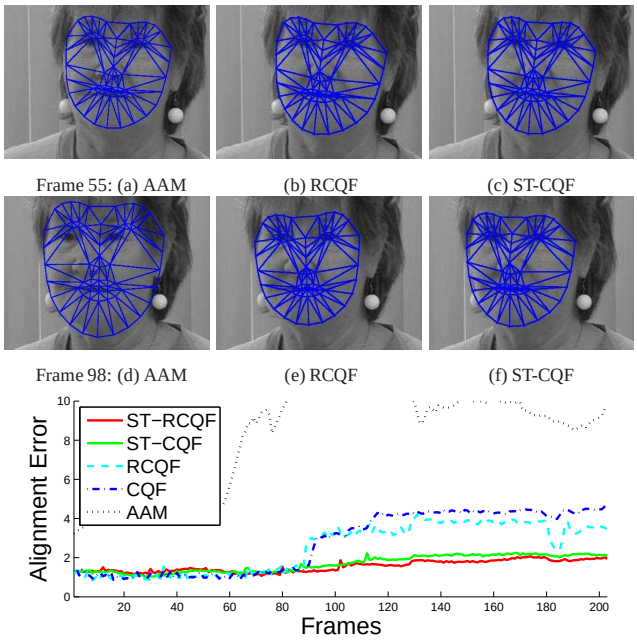
[15] L. Liang, F. Wen, Y. Xu, X. Tang, and H. Shum. Accurate face alignment using shape constrained Markov network. In *CVPR*, pages I: 1313–1319, 2006.

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



(g) Tracking Error Comparison

Figure 3. Examples of tracking performance on an unseen facial expression sequence. Since the MultiPIE face database [10] 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 1, 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.



(g) Tracking Error Comparison

Figure 4. 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 1, 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.

[17] 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.

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

[19] B.-J. Theobald, I. Matthews, and S. Baker. Evaluating error functions for robust active appearance models. In *International Conference on Automatic Face and Gesture Recognition*, pages 149–154, 2006.

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

[21] 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.

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

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

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

[25] 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.