# Robust Canonical Time Warping for the Alignment of Grossly Corrupted Sequences

Yannis Panagakis*, Mihalis A. Nicolaou*, Stefanos Zafeiriou*, and Maja Pantic*,†
*Department of Computing, Imperial College London, 180 Queens Gate, London SW7 2AZ, U.K.
†EEMCS, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands
{i.panagakis,mihalis,s.zafeiriou,m.pantic}@imperial.ac.uk

## Abstract

*Temporal alignment of human behaviour from visual data is a very challenging problem due to a numerous reasons, including possible large temporal scale differences, inter/intra subject variability and, more importantly, due to the presence of gross errors and outliers. Gross errors are often in abundance due to incorrect localization and tracking, presence of partial occlusion etc. Furthermore, such errors rarely follow a Gaussian distribution, which is the de-facto assumption in machine learning methods. In this paper, building on recent advances on rank minimization and compressive sensing, a novel, robust to gross errors temporal alignment method is proposed. While previous approaches combine the dynamic time warping (DTW) with low-dimensional projections that maximally correlate two sequences, we aim to learn two underlying projection matrices (one for each sequence), which not only maximally correlate the sequences but, at the same time, efficiently remove the possible corruptions in any datum in the sequences. The projections are obtained by minimizing the weighted sum of nuclear and $\ell_1$ norms, by solving a sequence of convex optimization problems, while the temporal alignment is found by applying the DTW in an alternating fashion. The superiority of the proposed method against the state-of-the-art time alignment methods, namely the canonical time warping and the generalized time warping, is indicated by the experimental results on both synthetic and real datasets.*

## 1. Introduction

Accurate temporal alignment of data sequences is a challenging problem raised in bioinformatics [14], speech processing [12, 19], and computer vision [8, 11, 24, 27, 25, 26], among many scientific disciplines. The problem is defined as finding the temporal coordinate transformation that brings two given data sequences into alignment in time. Some particular applications in computer vision include the alignment and the temporal segmentation of human motion [27, 26], the alignment of facial and motion capture data [25], the alignment of Kinect data [24], and view invariant action recognition [8, 11].

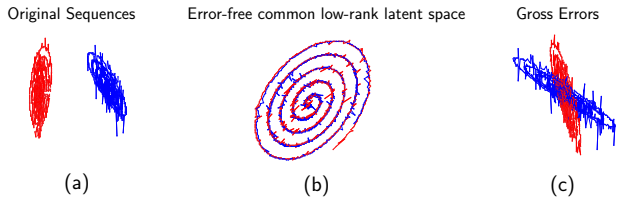


Figure 1. The RCTW applied on grossly-corrupted 3D data. (a) Original grossly-corrupted 3D data. (b) Alignment of the two data sequences onto an error-free common low-rank latent subspace which has been robustly estimated by the RCTW. (c) The removed gross errors.

The *dynamic time warping* (DTW) [19] aligns two sequences by minimizing the pairwise squared Euclidean distance via dynamic programming. Although, the DTW has been widely used for temporal alignment of data sequences, it has two main drawbacks: 1) the DTW fails under arbitrary affine transformations of one or both sequences and 2) cannot handle sequences with different dimensions. To alleviate the just mentioned drawbacks Zhou *et al.* proposed the *canonical time warping* (CTW) [25]. The CTW aligns two sequences in a common low-dimensional (or low-rank) latent subspace found by the canonical correlation analysis (CCA) [9]. The main limitation of the CTW is that, it is unable to handle sequences that lie on different manifolds. To this end, the *dynamic manifold temporal warping* (DMTW) [8] and the *manifold warping* (MW) [24] extend the CTW to handle more complex spatial transformations through man-

ifold learning. Since these methods rely on the DTW to find the temporal alignment, it is unclear how to adaptively constrain the temporal warping [26]. This drawback of the aforementioned DTW-based warping methods is addressed by the *isotonic cca* (ICCA) [21] and the *generalized time warping* (GTW) [26], where alternative constraints are imposed in order to guarantee monotonicity in the alignment space.

Despite the success of these methods in practise they are unable to uncover a common low-rank latent space for temporal alignment when high-dimensional data sequences are corrupted by gross non-Gaussian errors. Such errors are often occur in real video and motion capture data due to inaccurate tracking, illumination variations, partial occlusions, and gross pixel/angle corruptions. Indeed, it is known that the CCA-based alignment methods, discussed in the previous paragraph, are extremely fragile to the presence of gross corruptions [5]. This is a consequence of the conceptual similarity of the CCA with the principal component analysis [9].

In this paper, the *robust canonical time warping* (RCTW) is proposed for accurate temporal alignment of grossly corrupted high-dimensional data sequences. In particular, given two grossly corrupted high-dimensional data sequences, the RCTW aims to learn two low-rank projections which can efficiently remove the possible corruptions in the original noisy data sequences while simultaneously finding the temporal alignment that maximizes the spatial correlation between the error-free data sequences. In other words, the RCTW aligns the corrupted sequences in a error-free common low-rank latent subspace which is robustly estimated, even in the presence of gross errors. The projections are obtained by minimizing the weighted sum of nuclear and $\ell_1$ norms, by solving a sequence of convex optimization problems, while the temporal alignment is found by applying the DTW in an alternating fashion. An illustrative example of the working principle of the RCTW is shown in Fig. 1. Unlike the small Gaussian noise assumed in the CTW, the GTW, and the ICCA (due to the involvement of the CCA) [10], the RCTW can handle adequately the gross corruptions of large magnitude [1], provided that the corruptions are sparse enough (i.e., only a fraction of entries are corrupted). The RCTW model is mainly motivated by the success of *robust principal component analysis* (RPCA) [5] and *inductive RPCA* (IRPCA) [1] in gross error correction, and especially from the successful combination of rank minimization principles with spatial alignment [18].

The effectiveness of the RCTW in temporal alignment of grossly corrupted data sequences is assessed both visually and quantitatively by conducting 3 sets of experiments: 1) on the alignment of grossly corrupted synthetic data, 2) on the alignment of human walking in the presence of large occlusions, and 3) on the alignment of similar facial expressions made by two different individuals in the presence of noise spikes.

To summarize, the contributions of the paper are as follows:

- A novel method i.e., the RCTW, is proposed for accurate temporal alignment of high-dimensional data sequences despite large occlusions and corruptions.
- An efficient algorithm for the RCTW is derived by solving a sequence of convex problems. Each of the these convex problems is solved efficiently by employing first-order optimization techniques.
- Three different sets of experiments on synthetic and real video data validate that the proposed method accurately aligns grossly corrupted data sequences compared to state-of-the-art alignment methods, namely the CTW [25] and the GTW [26].

The paper is organized as follows. In Section 2, basic notation conventions are introduced. The DTW and the CTW are briefly reviewed in Section 3. The RCTW is developed in Section 4. Experimental results are presented in Section 5. Conclusions are drawn in Section 6.

## 2. Notations

Throughout the paper, matrices are denoted by uppercase boldface letters (e.g., $\mathbf{X}, \mathbf{Y}$), vectors are denoted by lowercase boldface letters (e.g., $\mathbf{x}$, $\mathbf{y}$), and scalars appear as either uppercase or lowercase letters (e.g., $T, d, i, \mu, \epsilon$). $\mathbf{I}$ denotes the identity matrix of compatible dimensions. $\mathbf{1}$ is a vector of ones. The $i$th column of $\mathbf{X}$ is denoted as $\mathbf{x}_i$ and the set of real numbers is denoted by $\mathbb{R}$.

A variety of norms will be used. For example, $\|\mathbf{X}\|_0$ is $\ell_0$ quasi-norm counting the number of nonzero entries in $\mathbf{X}$. The matrix $\ell_1$ norm is denoted by $\|\mathbf{X}\|_1 = \sum_i \sum_j |x_{ij}|$. $\|\mathbf{X}\|_F = \sqrt{\sum_i \sum_j x_{ij}^2} = \sqrt{\text{tr}(\mathbf{X}^T\mathbf{X})}$ is the Frobenius norm, where $\text{tr}(\cdot)$ denotes the trace of a square matrix. $\|\mathbf{x}\|_2$ denotes the $\ell_2$ norm. The nuclear norm of $\mathbf{X}$ is denoted by $\|\mathbf{X}\|_*$ and it is defined as the sum of its singular values.

## 3. Time Warping

To make the paper self-contained the DTW [19] and the CTW [25] are briefly reviewed.

### 3.1. Dynamic Time Warping

Given two data sequences $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\dots|\mathbf{x}_{T_x}] \in \mathbb{R}^{d\times T_x}$ and $\mathbf{Y} = [\mathbf{y}_1|\mathbf{y}_2|\dots|\mathbf{y}_{T_y}] \in \mathbb{R}^{d\times T_y}$, the DTW aligns the sequences by solving [19]:

$$\begin{aligned} \underset{\boldsymbol{\Delta}_x,\boldsymbol{\Delta}_y}{\text{argmin}} \ & \frac{1}{2}\|\mathbf{X}\boldsymbol{\Delta}_x - \mathbf{Y}\boldsymbol{\Delta}_y\|_F^2 \\ \text{s.t.} \ & \boldsymbol{\Delta}_x \in \{0,1\}^{T_x\times T}, \boldsymbol{\Delta}_y \in \{0,1\}^{T_y\times T}, \end{aligned} \tag{1}$$

where $\mathbf{\Delta}_x$ and $\mathbf{\Delta}_y$ are binary selection matrices encoding the alignment path. Although the number of possible alignments is exponential in $T_xT_y$, the DTW is able to recover the optimal alignment path in $\mathcal{O}(T_xT_y)$ by employing dynamic programming.

### 3.2. Canonical Time Warping

The CTW [25] incorporates CCA into the DTW, allowing the alignment of data sequences of different dimensions by projecting them into a common latent subspace found by CCA [9]. Furthermore, the CCA-based projections perform feature selection by reducing the dimensionality of the data to that of the common latent subspace, handling the irrelevant or possibly noisy attributes.

Formally, let $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\dots|\mathbf{x}_{T_x}] \in \mathbb{R}^{d_x \times T_x}$ and $\mathbf{Y} = [\mathbf{y}_1|\mathbf{y}_2|\dots|\mathbf{y}_{T_y}] \in \mathbb{R}^{d_y \times T_y}$ be two data sequences of different dimensionality (i.e., $d_x \neq d_y$), the CCA is incorporated into the DTW by solving [25]:

$$\begin{aligned}\underset{\mathbf{V}_x,\mathbf{V}_y,\mathbf{\Delta}_x,\mathbf{\Delta}_y}{\operatorname{argmin}} \quad & \frac{1}{2}\|\mathbf{V}_x\mathbf{X}\mathbf{\Delta}_x - \mathbf{V}_y\mathbf{Y}\mathbf{\Delta}_y\|_F^2 \\ \text{s.t.} \quad & \mathbf{X}\mathbf{\Delta}_x\mathbf{1} = \mathbf{0}, \;\; \mathbf{Y}\mathbf{\Delta}_y\mathbf{1} = \mathbf{0}, \\ & \mathbf{V}_x\mathbf{X}\mathbf{\Delta}_x\mathbf{\Delta}_x^T\mathbf{X}^T\mathbf{V}_x^T = \mathbf{I}, \\ & \mathbf{V}_y\mathbf{Y}\mathbf{\Delta}_y\mathbf{\Delta}_y^T\mathbf{Y}^T\mathbf{V}_y^T = \mathbf{I}, \\ & \mathbf{V}_x\mathbf{X}\mathbf{\Delta}_x\mathbf{\Delta}_y^T\mathbf{Y}^T\mathbf{V}_y^T = \mathbf{D}, \\ & \mathbf{\Delta}_x \in \{0,1\}^{T_x \times T}, \mathbf{\Delta}_y \in \{0,1\}^{T_y \times T}.\end{aligned} \tag{2}$$

$\mathbf{V}_x \in \mathbb{R}^{d' \times d_x}$ and $\mathbf{V}_y\mathbb{R}^{d' \times d_y}$ project $\mathbf{X}$ and $\mathbf{Y}$, respectively onto a common latent subspace of $d' \leq \min(d_x, d_y)$ dimensions, where the correlation between the data sequences is maximized. $\mathbf{D}$ is a diagonal matrix of compatible dimensions. The set of constraints in (2) is imposed in order to make the CTW translation, rotation, and scaling invariant. The solution of (2) is obtained by solving CCA and DTW in an alternating fashion.

## 4. Robust Canonical Time Warping

In this section, the alignment of high-dimensional data sequences in the presence of noise is investigated. Provided that the errors in the data sequences follow Gaussian distribution with small variance, the CCA is still able to uncover the common low-rank latent subspace and thus the CTW will accurately align two such noisy data sequences.

However, in real world conditions, the performance of the CCA and thus that of the CCA-based time warping methods (i.e., CTW and GTW) is limited since the CCA is not robust to gross corruptions. That is, the estimation of the common low-rank latent subspace found by the CCA could be far away from the underlying true common subspace in the presence of gross corruptions [1].

To this end, the RCTW is proposed as a *robust to gross errors* extension of the CTW. Consequently, the main aim of the RCTW is to learn two low-rank projections, which are able to uncover a common error-free low-rank subspace for the temporal alignment. Let $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\dots|\mathbf{x}_{T_x}] \in \mathbb{R}^{d \times T_x}$ and $\mathbf{Y} = [\mathbf{y}_1|\mathbf{y}_2|\dots|\mathbf{y}_{T_y}] \in \mathbb{R}^{d \times T_y}$ be the high-dimensional grossly corrupted data sequences to be aligned, $\mathbf{P}_x \in \mathbb{R}^{d \times d}$ and $\mathbf{P}_y \in \mathbb{R}^{d \times d}$ are the low-rank projections matrices, and $\mathbf{\Delta}_x \in \{0,1\}^{T_x \times T}$, $\mathbf{\Delta}_y \in \{0,1\}^{T_y \times T}$ encode the alignment path. Based on the desired low-rankness of the projections and the sparsity of the noise, the unknown matrices $\mathbf{P}_x, \mathbf{P}_y, \mathbf{\Delta}_x, \mathbf{\Delta}_y$ as well as the sparse distortion terms (i.e., $\mathbf{E}_x$ and $\mathbf{E}_y$) can be found by solving:

$$\begin{aligned}\underset{\substack{\mathbf{P}_x,\mathbf{P}_y,\mathbf{E}_x,\\ \mathbf{E}_y,\mathbf{\Delta}_x,\mathbf{\Delta}_y}}{\operatorname{argmin}} \quad & \operatorname{rank}(\mathbf{P}_x) + \operatorname{rank}(\mathbf{P}_y) + \lambda_x\|\mathbf{E}_x\|_0 + \lambda_y\|\mathbf{E}_y\|_0 \\ & + \frac{\mu}{2}\|\mathbf{P}_x\mathbf{X}\mathbf{\Delta}_x - \mathbf{P}_y\mathbf{Y}\mathbf{\Delta}_y\|_F^2 \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{P}_x\mathbf{X} + \mathbf{E}_x, \mathbf{Y} = \mathbf{P}_y\mathbf{Y} + \mathbf{E}_y, \\ & \mathbf{\Delta}_x \in \{0,1\}^{T_x \times T}, \mathbf{\Delta}_y \in \{0,1\}^{T_y \times T}.\end{aligned} \tag{3}$$

where $\lambda_x, \lambda_y, \mu$ are nonnegative parameters.

Problem (3) is difficult to solve due to the discrete nature of the rank function [23] and the $\ell_0$ norm [15]. A convex relaxation of (3) is obtained by replacing the rank function and the $\ell_0$ norm by their convex envelopes as follows, namely by the nuclear norm [7] and the $\ell_1$ norm [6], respectively as follows:

$$\begin{aligned}\underset{\substack{\mathbf{P}_x,\mathbf{P}_y,\mathbf{E}_x,\\ \mathbf{E}_y,\mathbf{\Delta}_x,\mathbf{\Delta}_y}}{\operatorname{argmin}} \quad & \|\mathbf{P}_x\|_* + \|\mathbf{P}_y\|_* + \lambda_x\|\mathbf{E}_x\|_1 + \lambda_y\|\mathbf{E}_y\|_1 \\ & + \frac{\mu}{2}\|\mathbf{P}_x\mathbf{X}\mathbf{\Delta}_x - \mathbf{P}_y\mathbf{Y}\mathbf{\Delta}_y\|_F^2 \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{P}_x\mathbf{X} + \mathbf{E}_x, \mathbf{Y} = \mathbf{P}_y\mathbf{Y} + \mathbf{E}_y, \\ & \mathbf{\Delta}_x \in \{0,1\}^{T_x \times T}, \mathbf{\Delta}_y \in \{0,1\}^{T_y \times T}.\end{aligned} \tag{4}$$

Problem (4) can be solved iteratively by employing the *linearized alternating directions method* (LADM) [13], which is a variant of the *alternating direction augmented Lagrange multiplier method* (ADM) [3]. That is, (4) is solved by minimizing the (partial) augmented Lagrangian function:

$$\begin{aligned}&\mathcal{L}(\mathbf{P}_x, \mathbf{P}_y, \mathbf{E}_x, \mathbf{E}_y, \mathbf{\Delta}_x, \mathbf{\Delta}_y, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2) \\ &= \|\mathbf{P}_x\|_* + \|\mathbf{P}_y\|_* + \lambda_x\|\mathbf{E}_x\|_y + \lambda_2\|\mathbf{E}_y\|_1 \\ &+ \frac{\mu}{2}\|\mathbf{P}_x\mathbf{X}\mathbf{\Delta}_x - \mathbf{P}_y\mathbf{Y}\mathbf{\Delta}_y\|_F^2 \\ &+ \operatorname{tr}\left(\mathbf{\Lambda}_1{}^T(\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x)\right) \\ &+ \operatorname{tr}\left(\mathbf{\Lambda}_2{}^T(\mathbf{Y} - \mathbf{P}_y\mathbf{Y} - \mathbf{E}_y)\right) \\ &+ \frac{\mu_x}{2}\|\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x\|_F^2 + \frac{\mu_y}{2}\|\mathbf{Y} - \mathbf{P}_y\mathbf{Y} - \mathbf{E}_y\|_F^2 \\ &\text{s.t.} \;\; \mathbf{\Delta}_x \in \{0,1\}^{T_x \times T}, \mathbf{\Delta}_y \in \{0,1\}^{T_y \times T},\end{aligned} \tag{5}$$

where $\mathbf{\Lambda}_1, \mathbf{\Lambda}_2$ are the Lagrange multipliers for the equality constraints in (4) and $\mu_x, \mu_y$ are nonnegative penalty parameters. By employing the LADM, (5) is minimized with respect to each variable in an alternating fashion and finally the Lagrange multipliers are updated at each iteration as outlined in Algorithm 1. The derivation of Algorithm 1 is provided next.

If only $\mathbf{P}_x$ is varying and all the other variables are kept fixed, we simplify (5) writing $\mathcal{L}(\mathbf{P}_x)$ instead of $\mathcal{L}(\mathbf{P}_x, \mathbf{P}_y, \mathbf{E}_x, \mathbf{E}_y, \mathbf{\Delta}_x, \mathbf{\Delta}_y, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$. Let $t$ denote the iteration index, given $\mathbf{P}_{x[t]}$, $\mathbf{P}_{y[t]}$, $\mathbf{E}_{x[t]}$, $\mathbf{E}_{y[t]}$, $\mathbf{\Delta}_{x[t]}$, $\mathbf{\Delta}_{y[t]}$,$\mathbf{\Lambda}_{1[t]}$, and $\mathbf{\Lambda}_{2[t]}$, the iterative scheme of LADM for (5) reads as follows:

$$\mathbf{P}_{x[t+1]} = \underset{\mathbf{P}_{x[t]}}{\operatorname{argmin}} \mathcal{L}(\mathbf{P}_{x[t]}) \tag{6}$$

$$\mathbf{E}_{x[t+1]} = \underset{\mathbf{E}_{x[t]}}{\operatorname{argmin}} \mathcal{L}(\mathbf{E}_{x[t]}) \tag{7}$$

$$\mathbf{P}_{y[t+1]} = \underset{\mathbf{P}_{y[t]}}{\operatorname{argmin}} \mathcal{L}(\mathbf{P}_{y[t]}) \tag{8}$$

$$\mathbf{E}_{y[t+1]} = \underset{\mathbf{E}_{y[t]}}{\operatorname{argmin}} \mathcal{L}(\mathbf{E}_{y[t]}) \tag{9}$$

$$(\mathbf{\Delta}_{x[t+1]}, \mathbf{\Delta}_{y[t+1]}) = \underset{\mathbf{\Delta}_{x[t]}, \mathbf{\Delta}_{y[t]}}{\operatorname{argmin}} \mathcal{L}(\mathbf{\Delta}_{x[t]}, \mathbf{\Delta}_{y[t]}) \tag{10}$$

*Solving subproblems (6) and (8).* By fixing the other variables, subproblem (6) is reduced to

$$\begin{aligned}&\underset{\mathbf{P}_{x[t]}}{\operatorname{argmin}} \|\mathbf{P}_x\|_* + \frac{\mu}{2}\|\mathbf{P}_x\mathbf{X}\mathbf{\Delta}_x - \mathbf{P}_y\mathbf{Y}\mathbf{\Delta}_y\|_F^2 \\ &+ \operatorname{tr}\left(\mathbf{\Lambda}_1{}^T(\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x)\right) + \frac{\mu_x}{2}\|\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x\|_F^2.\end{aligned} \tag{11}$$

Although the standard procedure for solving nuclear norm regularized least squares problems is the *singular value thresholding* operator [4], it cannot be directly applied in case of (11), due to the existence of the second term (i.e., $\frac{\mu}{2}\|\mathbf{P}_x\mathbf{X}\mathbf{\Delta}_x - \mathbf{P}_y\mathbf{Y}\mathbf{\Delta}_y\|_F^2$). To this end, following [13], the differentiable terms in (11) i.e., the function $f(\mathbf{P}_x) = \frac{\mu}{2}\|\mathbf{P}_x\mathbf{X}\mathbf{\Delta}_x - \mathbf{P}_y\mathbf{Y}\mathbf{\Delta}_y\|_F^2 + \operatorname{tr}\left(\mathbf{\Lambda}_1{}^T(\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x)\right) + \frac{\mu_x}{2}\|\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x\|_F^2$ is linearly approximated with respect to $\mathbf{P}_x$ at $\mathbf{P}_{x[t]}$ as follows:

$$\begin{aligned}f(\mathbf{P}_x) \approx\ & f(\mathbf{P}_{x[t]}) + \operatorname{tr}\left((\mathbf{P}_x - \mathbf{P}_{x[t]})^T\nabla f(\mathbf{P}_{x[t]})\right) \\ &+ \frac{\mu_x\eta_x}{2}\|\mathbf{P}_x - \mathbf{P}_{x[t]}\|_F^2,\end{aligned} \tag{12}$$

where, $\eta_x$ is a proximal parameter. The gradient of $f(\mathbf{P}_{x[t]})$ with respect to $\mathbf{P}_{x[t]}$ is given by:

$$\begin{aligned}&\nabla f(\mathbf{P}_{x[t]}) = \mu_x(\mathbf{P}_{x[t]}\mathbf{X}\mathbf{X}^T + \mathbf{E}_{x[t]}\mathbf{X}^T - \mathbf{X}\mathbf{X}^T) \\ &+ \mu(\mathbf{P}_{x[t]}\mathbf{X}\mathbf{\Delta}_{x[t]}\mathbf{\Delta}_{x[t]}^T\mathbf{X}^T - \mathbf{P}_{y[t]}\mathbf{Y}\mathbf{\Delta}_{y[t]}\mathbf{\Delta}_{x[t]}^T\mathbf{X}^T) \\ &- \mathbf{\Lambda}_{1[t]}\mathbf{X}^T.\end{aligned} \tag{13}$$

Consequently, an approximate solution of (11) can be obtained as follows:

$$\begin{aligned}&\mathbf{P}_{x[t+1]} \approx \underset{\mathbf{P}_x}{\operatorname{argmin}} \|\mathbf{P}_x\|_* + f(\mathbf{P}_{x[t]}) \\ &+ \operatorname{tr}\left((\mathbf{P}_x - \mathbf{P}_{x[t]})^T\nabla f(\mathbf{P}_{x[t]})\right) + \frac{\mu_x\eta_x}{2}\|\mathbf{P}_x - \mathbf{P}_{x[t]}\|_F^2 \\ &= \underset{\mathbf{P}_x}{\operatorname{argmin}} \|\mathbf{P}_x\|_* + \frac{\mu_x\eta_x}{2}\|\mathbf{P}_x - (\mathbf{P}_{[t]} - \frac{1}{\mu_x\eta_x}\nabla f(\mathbf{P}_{x[t]})\|_F^2 \\ &= \mathcal{D}_{\frac{1}{\mu_x\eta_x}}\left[\mathbf{P}_{x[t]} - \frac{1}{\mu_x\eta_x}\nabla f(\mathbf{P}_{x[t]})\right].\end{aligned} \tag{14}$$

The singular value thresholding operator defined for any matrix $\mathbf{Q}$ as [4]: $\mathcal{D}_\tau[\mathbf{Q}] = \mathbf{U}\mathcal{S}_\tau\mathbf{V}^T$ with $\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ being the singular value decomposition and $\mathcal{S}_\tau[q] = \operatorname{sgn}(q)\max(|q| - \tau, 0)$ is the shrinkage operator [5], which can be extended to matrices by applying it element-wise.

The solution of (8) in analogy with (6) is given by

$$\mathbf{P}_{y[t+1]} = \mathcal{D}_{\frac{1}{\mu_y\eta_y}}\left[\mathbf{P}_{y[t]} - \frac{1}{\mu_y\eta_y}\nabla f(\mathbf{P}_{y[t]})\right], \tag{15}$$

where $\nabla f(\mathbf{P}_{y[t]}) = \mu_x(\mathbf{P}_{y[t]}\mathbf{Y}\mathbf{Y}^T + \mathbf{E}_{y[t]}\mathbf{Y}^T - \mathbf{Y}\mathbf{Y}^T) + \mu(\mathbf{P}_{y[t]}\mathbf{Y}\mathbf{\Delta}_{y[t]}\mathbf{\Delta}_{y[t]}^T\mathbf{Y}^T - \mathbf{P}_{x[t]}\mathbf{X}\mathbf{\Delta}_{x[t]}\mathbf{\Delta}_{y[t]}^T\mathbf{Y}^T) - \mathbf{\Lambda}_{2[t]}\mathbf{Y}^T$.

*Solving subproblems (7) and (9).* By fixing the other variables, subproblem (7) is reduced to

$$\begin{aligned}\underset{\mathbf{E}_{x[t]}}{\operatorname{argmin}} &\lambda_x\|\mathbf{E}_x\|_1 + \operatorname{tr}\left(\mathbf{\Lambda}_1{}^T(\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x)\right) \\ &+ \frac{\mu_x}{2}\|\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x\|_F^2.\end{aligned} \tag{16}$$

The subgradient of (16) provides a closed-form solution for $\mathbf{E}_{x[t+1]}$ by employing the shrinkage operator:

$$\mathbf{E}_{x[t+1]} = \mathcal{S}_{\frac{\lambda_x}{\mu_x}}[\mathbf{X} - \mathbf{P}_{x[t+1]}\mathbf{X} + \frac{1}{\mu_x}\mathbf{\Lambda}_{1[t]}]. \tag{17}$$

In a similar manner to (7), the solution of (9) is given by:

$$\mathbf{E}_{y[t+1]} = \mathcal{S}_{\frac{\lambda_y}{\mu_y}}[\mathbf{Y} - \mathbf{P}_{y[t+1]}\mathbf{Y} + \frac{1}{\mu_y}\mathbf{\Lambda}_{2[t]}]. \tag{18}$$

*Solving (10).* Subproblem (10) is solved by applying the DTW on the clean latent spaces defined by $\mathbf{P}_{x[t+1]}\mathbf{X}, \mathbf{P}_{y[t+1]}\mathbf{Y}$. Thus the warping matrices are obtained as follows:

$$[\mathbf{\Delta}_{x[t+1]}, \mathbf{\Delta}_{y[t+1]}] = \operatorname{DTW}(\mathbf{P}_{x[t+1]}\mathbf{X}, \mathbf{P}_{y[t+1]}\mathbf{Y}). \tag{19}$$

The Algorithm 1 terminates when the following criteria are satisfied [13]:

$$\begin{aligned}\max\big(&\frac{\|\mathbf{X} - \mathbf{P}_{x[t+1]}\mathbf{X} - \mathbf{E}_{x[t+1]}\|_F}{\|\mathbf{X}\|_F}, \\ &\frac{\|\mathbf{Y} - \mathbf{P}_{y[t+1]}\mathbf{Y} - \mathbf{E}_{y[t+1]}\|_F}{\|\mathbf{Y}\|_F}\big) < \epsilon_1,\end{aligned} \tag{20}$$

**Algorithm 1** Solving (5) by the LADM method.

**Input:** Data sequences: $\mathbf{X} \in \mathbb{R}^{d \times T_x}$ and $\mathbf{Y} = \in \mathbb{R}^{d \times T_y}$, parameters: $\lambda_x = 1/\sqrt{\max(d, T_x)}$, $\lambda_y = 1/\sqrt{\max(d, T_y)}$.
**Output:** The projection matrices: $\mathbf{P}_x, \mathbf{P}_y$, the warping matrices $\mathbf{\Delta}_x, \mathbf{\Delta}_y$, and the error matrices $\mathbf{E}_x, \mathbf{E}_y$.

1: Initialize: Set $\mathbf{P}_{x[0]}, \mathbf{P}_{y[0]}, \mathbf{E}_{x[0]}$, and $\mathbf{E}_{y[0]}$ to zero matrices of compatible dimensions. Initialize $\mathbf{\Delta}_{x[0]}$ and $\mathbf{\Delta}_{y[0]}$ by the DTW. $t = 0$ $\mu_{[0]} = \mu_{x[0]} = \mu_{y[0]} = 10^{-6}$, $\rho = 1.9$, $\eta_x = 1.02\sigma_x^2$, $\eta_y = 1.02\sigma_y^2$, where $\sigma_x$, $\sigma_y$ are the largest singular values of $\mathbf{X}$ and $\mathbf{Y}$, respectively. $\epsilon_1 = 10^{-4}, \epsilon_2 = 10^{-5}$.
2: **while** not converged **do**
3: Fix the other variables, and update $\mathbf{P}_{x[t+1]}$ by: $\mathbf{P}_{x[t+1]} \leftarrow \mathcal{D}_{\frac{1}{\mu_{x[t]}\eta_x}}[\mathbf{P}_{x[t]} - 1/(\mu_{x[t]} \cdot \eta_x)\nabla f(\mathbf{P}_{x[t]})]$.
4: Fix the other variables, and update $\mathbf{E}_{x[t+1]}$ by: $\mathbf{E}_{x[t+1]} \leftarrow \mathcal{S}_{\frac{\lambda_1}{\mu_{x[t]}}}[\mathbf{X} - \mathbf{P}_{x[t+1]}\mathbf{X} + \frac{1}{\mu_{x[t]}}\mathbf{\Lambda}_{1[t]}]$.
5: Fix the other variables, and update $\mathbf{P}_{y[t+1]}$ by: $\mathbf{P}_{y[t+1]} \leftarrow \mathcal{D}_{\frac{1}{\mu_{y[t]}\eta_y}}[\mathbf{P}_{y[t]} - 1/(\mu_{y[t]} \cdot \eta_y)\nabla f(\mathbf{P}_{y[t]})]$.
6: Fix the other variables, and update $\mathbf{E}_{y[t+1]}$ by: $\mathbf{E}_{y[t+1]} \leftarrow \mathcal{S}_{\frac{\lambda_2}{\mu_{y[t]}}}[\mathbf{Y} - \mathbf{P}_{y[t+1]}\mathbf{Y} + \frac{1}{\mu_{y[t]}}\mathbf{\Lambda}_{2[t]}]$.
7: Fix the other variables, and update the warping paths $\mathbf{\Delta}_{x[t+1]}, \mathbf{\Delta}_{y[t+1]}$ by: $[\mathbf{\Delta}_{x[t+1]}, \mathbf{\Delta}_{y[t+1]}] \leftarrow \text{DTW}(\mathbf{P}_{x[t+1]}\mathbf{X}, \mathbf{P}_{y[t+1]}\mathbf{Y})$.
8: Update the Lagrange multipliers by: $\mathbf{\Lambda}_{1[t+1]} \leftarrow \mathbf{\Lambda}_{1[t]} + \mu_{x[t]}(\mathbf{X} - \mathbf{P}_{x[t+1]}\mathbf{X} - \mathbf{E}_{x[t+1]})$. $\mathbf{\Lambda}_{2[t+1]} \leftarrow \mathbf{\Lambda}_{2[t]} + \mu_{y[t]}(\mathbf{Y} - \mathbf{P}_{y[t+1]}\mathbf{Y} - \mathbf{E}_{y[t+1]})$.
9: Update $\mu_{x[t+1]}$ by:
10: **if** $\mu_{x[t]}\|\mathbf{P}_{x[t+1]} - \mathbf{P}_{x[t]}\|_F / \|\mathbf{X}\|_F \leq \epsilon_2$ **then**
11: $\mu_{x[t+1]} \leftarrow \min(\rho \cdot \mu_{x[t]}, 10^6)$.
12: **end if**
13: **if** $\mu_{y[t]}\|\mathbf{P}_{y[t+1]} - \mathbf{P}_{y[t]}\|_F / \|\mathbf{Y}\|_F \leq \epsilon_2$ **then**
14: $\mu_{y[t+1]} \leftarrow \min(\rho \cdot \mu_{y[t]}, 10^6)$.
15: **end if**
16: Update $\mu_{[t+1]}$ by: $\mu_{[t+1]} \leftarrow \min(\mu_{x[t+1]}, \mu_{y[t+1]})$
17: Check convergence conditions in (21) and (20).
18: $t \leftarrow t + 1$.
19: **end while**

and

$$\max\Big(\frac{\|\mathbf{P}_{x[t+1]} - \mathbf{P}_{x[t]}\|_F}{\|\mathbf{X}\|_F}, \frac{\|\mathbf{P}_{y[t+1]} - \mathbf{P}_{y[t]}\|_F}{\|\mathbf{Y}\|_F}, \frac{\|\mathbf{E}_{x[t+1]} - \mathbf{E}_{x[t]}\|_F}{\|\mathbf{X}\|_F}, \frac{\|\mathbf{E}_{y[t+1]} - \mathbf{E}_{y[t]}\|_F}{\|\mathbf{Y}\|_F}\Big) < \epsilon_2. \quad (21)$$

The dominant cost of each iteration in Algorithm 1 is the computation the singular value thresholding operator (i.e., Step 3 and Step 5). Thus, the complexity of each iteration is $\mathcal{O}(d^2 \cdot T)$. Regarding the convergence of Algorithm 1, there is no established convergence proof of the ADM for more than two blocks of variables [3, 18]. Nevertheless, weak convergence results can be derived if the block of variables is assumed to be bounded (e.g., Proposition 2.2 in [22]). However, the application of ADM in optimization problems with more than two blocks of variables (e.g., [1, 18]) yields algorithms whose convergence is empirically guaranteed. This can be attributed to the convexity of (5) with respect to all the blocks of variables.

If the dimensions of the data sequence are different i.e., $\mathbf{X} \in \mathbb{R}^{d_x \times T_x}$ and $\mathbf{Y} \in \mathbb{R}^{d_y \times T_y}$ with $d_y \neq d_x$, then the dimensionality of the largest sequence can be reduced to that of the smallest by a random projection matrix drawn from a normal zero-mean distribution. Such a random projection matrix provides with high probability a *stable embedding* [2] preserving the Euclidean distances between all vectors in the original space in the feature space of reduced dimensions. Furthermore, if both data sequences are high-dimensional such as videos, random projections could be applied to both of the for computational tractability.

## 5. Experimental Evaluation

In this section, the performance of the RCTW in temporal alignment is assessed by conducting experiments on both synthetic (Subsection 5.1) and real data (Subsection 5.2 and 5.3), contaminated by gross errors. Performance comparisons are made against the state-of-the-art temporal alignment methods, namely the CTW [25] and the GTW [26]. The alignment error is evaluated by employing the following metric [26]:

$$\text{Err} = \frac{\text{dist}(\mathbf{\Pi}^*, \hat{\mathbf{\Pi}}) + \text{dist}(\hat{\mathbf{\Pi}}, \mathbf{\Pi}^*)}{m^* + \hat{m}},$$
$$\text{dist}(\mathbf{\Pi}_1, \mathbf{\Pi}_2) = \sum_{i=1}^{m_1} \min(\{\|\pi_1^{(i)} - \pi_2^{(j)}\|_2\})_{j=1}^{m_2}), \quad (22)$$

where $m^*$ is the length of $\mathbf{\Pi}^*$ and $\hat{m}$ is the length $\hat{\mathbf{\Pi}}$.

### 5.1. Synthetic Data

For the synthetic experiments a similar setting to [25] was employed. That is, a set of 3D spirals data sequences were generated as follows: $\mathbf{X} = \mathbf{S}_x\mathbf{Z}\mathbf{T}_x \in \mathbb{R}^{3 \times T_x}$, $\mathbf{Y} = \mathbf{S}_y\mathbf{Z}\mathbf{T}_y \in \mathbb{R}^{3 \times T_y}$, where $\mathbf{Z} \in \mathbb{R}^{3 \times T}$ is the true latent data sequence. $\mathbf{S}_x, \mathbf{S}_y \in \mathbb{R}^{3 \times 3}$ and $\mathbf{T}_x \in \mathbb{R}^{T_x \times T}, \mathbf{T}_y \in \mathbb{R}^{T_y \times T}$ are random spatial and temporal warping matrices, respectively. Next, both $\mathbf{X}$ and $\mathbf{Y}$ are corrupted by adding gross non-gaussian noise to a percentage of samples (i.e., columns of $\mathbf{X}$ and $\mathbf{Y}$) ranging from 5 to 55%.

An example of the noisy synthetic data alignment obtained by the CTW, the GTW, and the proposed RCTW is depicted in Fig. 2. Clearly, the RCTW smooths the noise of

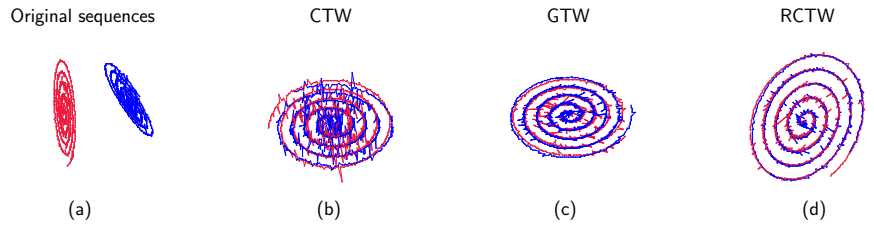


Figure 2. Alignment of synthetic data; 30% of the samples of each sequence have been contaminated by gross errors. (a) Initial noisy data sequences. The alignment achieved by (b) the CTW, (c) the GTW, and (d) the RCTW.

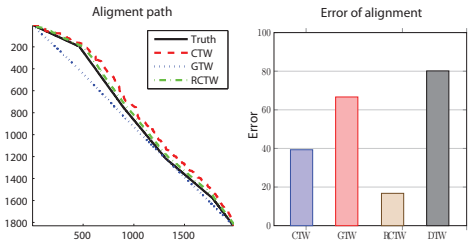


Figure 3. Comparison of the performance of the CTW, the GTW, and the RCTW on synthetic data alignment. The mean alignment path (left) and the mean alignment error (right) obtained by the CTW, the GTW, and the RCTW (left) by applying 50 different random spatial and temporal transformations on the latent data sequence $\mathbf{Z}$.

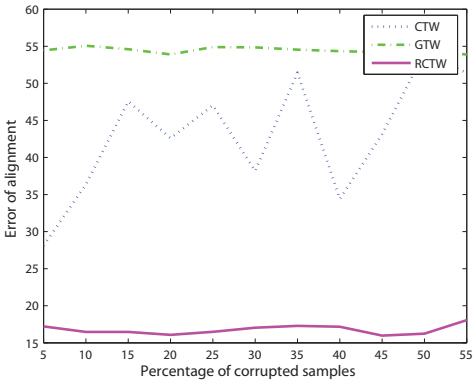


Figure 4. Mean alignment error obtained by the CTW, the GTW, and the RCTW, as a function of the percentage of corrupted samples on synthetic data sequences.

the initial data sequencers, yielding a better alignment than the CTW and the GTW.

In Fig. 3 we present averaged results on 50 data sequences, where the latent data sequence $\mathbf{Z}$ is perturbed by 50 different random spatial and temporal transformations. The mean alignment error of the compared techniques is presented in Fig. 4. It is clear from both figures that the RCTW outperforms the compared approaches, exhibiting a stable and low path alignment error.

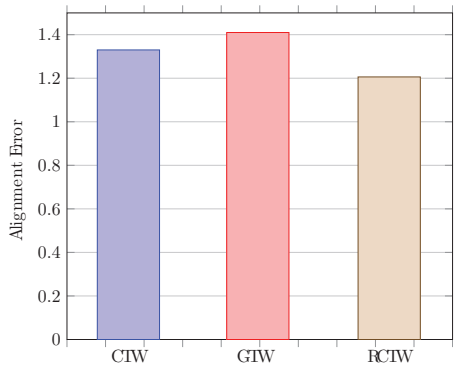


Figure 5. Mean alignment error obtained by the CTW, the GTW and the RCTW on human walking sequences by the KTH.

### 5.2. Real Data I: Temporal Alignment of Human Walking

In this set of experiments, the performance of the RCTW in alignment of human actions is assessed by conducting experiments on the KTH database [20]. To this end, 25 pairs of sequences consisting of videos performing the same action (walking) were randomly selected. Variations within the pairs appear in clothing, background or view angle. To make the experiment more challenging, we occlude 30% of each frame. In Fig. 5 the mean alignment error obtained by the CTW, the GTW and the RCTW on corrupted human walking sequences is depicted. Clearly, the RCTW outperforms the CTW and the GTW with respect to alignment error. An illustrative example of aligning occluded human walking sequences with the RCTW is depicted in Fig. 6. It can be observed that the occlusions have been removed.

### 5.3. Real Data II: Temporal Action Unit Alignment

The MMI dataset [16] has been employed in order to assess the performance of the RCTW on the temporal alignment of facial expressions. The MMI database [16] consists of more than 300 videos which have been annotated in terms of *action units* (AUs). In particular, each video contains frame-by-frame annotations of each action unit activated covering all temporal phases (i.e., neutral, onset, apex, offset) of each AU. We use a subset of the database with ap-

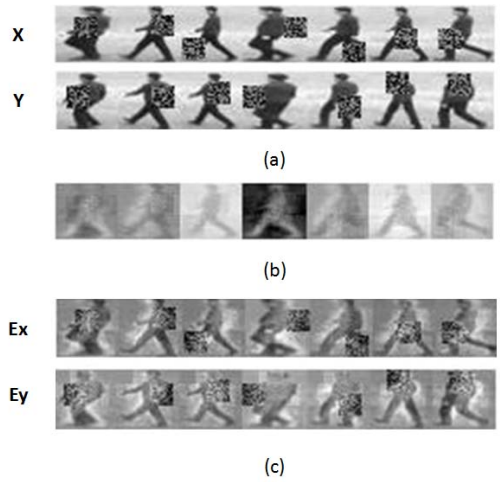


Figure 6. Alignment of occluded human walking sequences obtained by the RCTW. (a) The initial occluded walking sequences i.e., $\mathbf{X}$, $\mathbf{Y}$. (b) Aligned sequences onto the error-free latent common space which has been robustly estimated by the RCTW. (c) Magnitude of the recovered gross errors.

proximately $50$ pairs of videos of $8$ different subjects where action unit $12$ is activated.

The experiment proceeds as follows. Firstly, we extract a set of $20$ facial points using a person independent tracker presented in [17]. We use $8$ 2D points ($16$ dimensional feature vector) which refer to the lower face. Subsequently, we corrupt the facial features with sparse spike noise in order to evaluate the robustness of the compared algorithms. In particular, we draw values from a random normal distribution and add uniformly to $5\%$ of the frames of each video. This type of noise is common when using detection-based trackers, in which case a point can be misdirected for several frames.

Results are presented in Fig. 7. The error we used is the percentage of misaligned frames for each pair of videos, normalised per frame (i.e. divided by the aligned video length). We present results on average (for the entire video, Fig. 7(a)) and results regarding the apex (which is the 'peak' of the expression, Fig. 7(b)). In the presented results, the number of features corrupted by noise increases to $4$ out of $8$ (which essentially means that $50\%$ of our features are corrupted by noise). It is clear from the results that the RCTW can outperform both the CTW and the GTW in this scenario, maintaining relatively low error even when heavily increasing the presence of noise.

## 6. Conclusions

By exploiting recent advances on matrix rank minimization and compressive sensing we proposed the first method which simultaneously discovers a subspace, in which two sequences maximally correlate, and in the same time removes possibly gross errors from the data. The proposed method (i.e., the RCTW) outperforms the state-of-the-art techniques in temporal alignment of data sequences in the presence of gross errors.

### Acknowledgements

This work has been funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The work of Mihalis Nicolaou is funded by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 288235 (FROG).

## References

[1] B. Bao, G. Liu, C. Xu, and Y. S. Inductive robust principal component analysis. *IEEE Trans. Image Processing*, 21(8):3794–3800, 2012.

[2] R. Baraniuk, V. Cevher, and M. Wakin. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proceedings of the IEEE*, 98(6):959–971, 2010.

[3] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, Belmont, MA, 2nd edition, 1996.

[4] J. F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal Optimization*, 2(2):569–592, 2009.

[5] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58(3):1–37, 2011.

[6] D. Donoho. For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59(7):907–934, 2006.

[7] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Dept. Electrical Engineering, Stanford University, CA, USA, 2002.

[8] D. Gong and G. Medioni. Dynamic manifold warping for view invariant action recognition. In *Proc. 13th IEEE Int. Conf. Computer Vision*, pages 571–578, 2011.

[9] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[10] D. Huang, R. S. Cabral, and F. De la Torre. Robust regression. In *Proc. 12th European Conf. Computer Vision*, pages 616–630, 2012.

[11] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez. View-independent action recognition from temporal self-similarities. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(1):172–185, 2011.

[12] B. King, P. Smaragdis, and J. Mysore. Noise-robust dynamic time warping using PLCA features. In *Proc. 2012 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pages 1973–1976, 2012.

[13] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation.

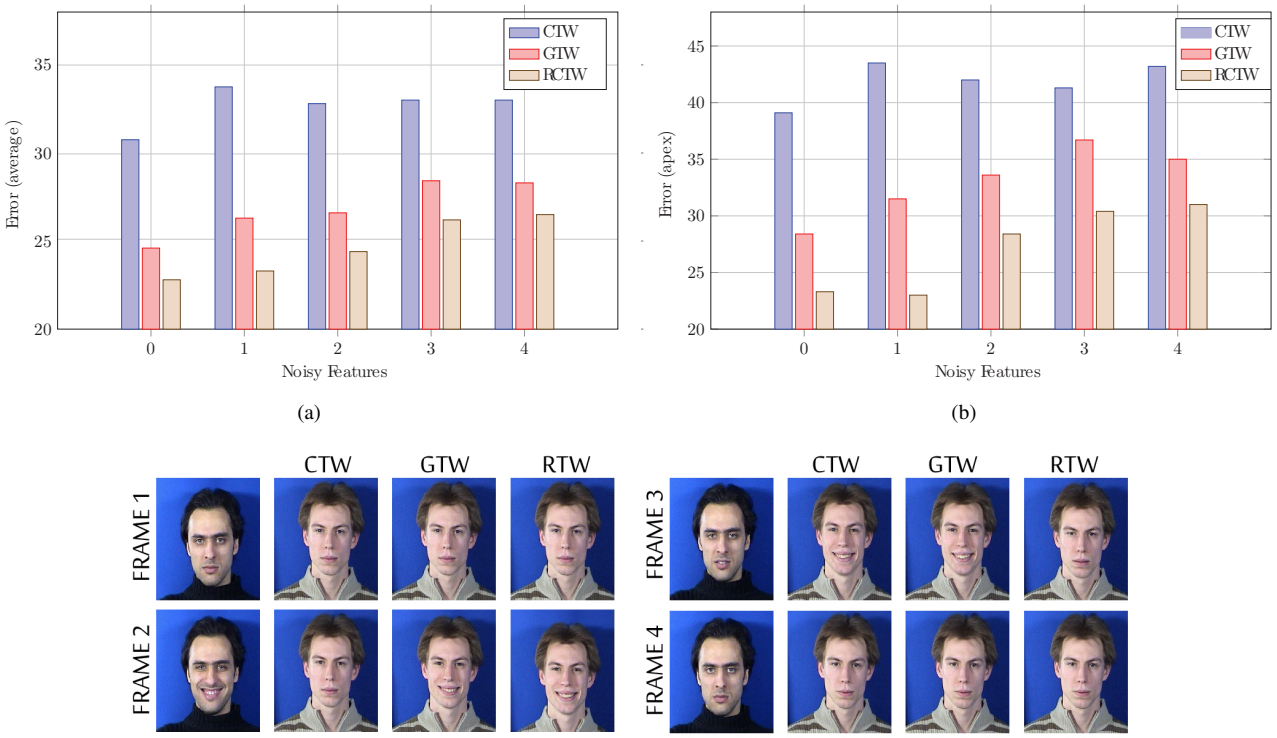


Figure 7. Action Unit alignment comparing the RCTW, the CTW, and the GTW. (a) Average error, (b) error for apex phase, (c) example video.


In *Proc. 2011 Neural Information Processing Systems Conf.*, pages 612–620, Granada, Spain, 2011.

[14] J. Listgarten, R. Neal, S. Roweis, and A. Emili. Multiple alignment of continuous time series. In *Advances in Neural Information Processing Systems*, volume 17, 2005.

[15] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

[16] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proc. 2005 IEEE Int. Conf. Multimedia and Expo*, Amsterdam, The Netherlands, 2005.

[17] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *Proc. 2004 IEEE Int. Conf. Automatic Face and Gesture Recognition*, pages 97–102, 2004.

[18] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34:2233–2246, 2012.

[19] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing*, (1):43–49, 1978.

[20] C. Schuldt, I. Laptev, and B. C. Recognizing human actions: A local SVM approach. In *Proc. 17th Int. Conf. Pattern Recognition*, pages 32–36, Washington, DC, USA, 2004.

[21] S. Shariat and V. Pavlovic. Isotonic CCA for sequence alignment and activity recognition. pages 2572–2578, 2011.

[22] Y. Shen, Z. Wen, and Y. Zhang. Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optimization Methods and Software*, pages 1–25, 2012.

[23] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

[24] H. Vu, C. Carey, and S. Mahadevan. Manifold warping: Manifold alignment over time. In *Proc. 26th Conference on Artificial Intelligence*, 2012.

[25] F. Zhou and F. De la Torre. Canonical time warping for alignment of human behavior. In *Proc. 2009 Advances in Neural Information Processing Systems*, pages 2286–2294. 2009.

[26] F. Zhou and F. De la Torre. Generalized time warping for multi-modal alignment of human motion. In *Proc. 2012 IEEE Conf. Computer Vision and Pattern Recognition*, pages 1282–1289, 2012.

[27] F. Zhou, F. De la Torre Torre, and J. K. Hodgins. Aligned cluster analysis for temporal segmentation of human motion. In *Proc. 2008 IEEE Int. Conf. Automatic Face and Gestures Recognition*, 2008.