Joint Unsupervised Deformable Spatio-Temporal Alignment of Sequences

Lazaros Zafeiriou* Epameinondas Antonakos* Stefanos Zafeiriou*‡
Maja Pantic*†
*Imperial College London, UK †University of Twente, The Netherlands
‡Center for Machine Vision and Signal Analysis, University of Oulu, Finland
*{l.zafeiriou12, e.antonakos, s.zafeiriou, m.pantic}@imperial.ac.uk, †PanticM@cs.utwente.nl

Abstract

Typically, the problems of spatial and temporal alignment of sequences are considered disjoint. That is, in order to align two sequences, a methodology that (non)-rigidly aligns the images is first applied, followed by temporal alignment of the obtained aligned images. In this paper, we propose the first, to the best of our knowledge, methodology that can jointly spatio-temporally align two sequences, which display highly deformable texture-varying objects. We show that by treating the problems of deformable spatial and temporal alignment jointly, we achieve better results than considering the problems independent. Furthermore, we show that deformable spatio-temporal alignment of faces can be performed in an unsupervised manner (i.e., without employing face trackers or building person-specific deformable models).

1. Introduction

Temporal and spatial alignment are two very well-studied fields in various disciplines, including computer vision and machine learning [34, 14, 35, 3, 16]. Temporal alignment is the first step towards analysis and synthesis of human and animal motion, temporal clustering of sequences and behaviour segmentation [34, 14, 35, 19, 30, 17, 36]. Spatial image alignment is among the main computer vision topics [3, 16, 1]. It is usually the first step towards many pattern matching applications such as face and facial expression recognition, object detection etc. [23, 24, 6]. It is also the first step towards temporal alignment of sequences [34, 35, 19, 30, 17].

Typically, temporal and spatial alignment are treated as two disjoint problems. Thus, they are solved separately, usually by employing very different methodologies. This is more evident in the task of spatio-temporal alignment of sequences that contain deformable objects. For example, the typical framework for temporal alignment of two sequences displaying objects that undergo non-rigid deformations, e.g. a facial expression, is the following [34, 35, 19, 30, 17]:

1. The first step is to apply a statistical facial deformable model (generic or person-specific) which aligns the images and/or localizes a consistent set of facial landmarks. Some examples of such state-of-the-art models are [28, 16]. Even though such deformable models demonstrate great capabilities, they require either thousands of manually annotated facial samples captured under various recording conditions (generic models) or the manual annotation of a set of frames in each and every video that is analysed (person-specific models). However, such extended manual annotation is a laborious and labour intensive procedure [22, 2].

2. The second step is to use the acquired densely aligned images or the localized landmarks to perform temporal alignment. However, one of the main challenges in aligning such visual data is their high dimensionality. This is the reason why various recently proposed methods perform temporal alignment by joint feature extraction and dimensionality reduction [34, 30, 17].

Joint spatio-temporal alignment is more advantageous than spatial alignment, since the spatial ambiguities that may be present can be resolved. The alignment accuracy can also be improved, because all the available information is exploited. Despite those advantages, joint spatio-temporal alignment has received limited attention, mainly due to the difficulty in designing such frameworks [8, 12]. The methods that have been proposed typically assume rigid spatial and temporal motion models (i.e., affine-like) [8]. Also, the video sequences display different views of the same dynamic scene [8, 12]. Hence, such methods are not suitable for the task of spatio-temporal alignment of sequences with deformable objects, such as faces.

To the best of our knowledge, no method has been proposed that is able to perform deformable joint spatio-temporal alignment of sequences that contain texture-

1063-6919/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPR.2016.368

varying deformable objects (e.g., faces). The existing methods for aligning such sequences usually require hours of manual annotation in order to first develop models that are able to extract deformations (commonly described by a set of sparse tracked landmarks), and then align the extracted deformations [34, 30, 17]. An additional advantage of methodologies that can jointly spatio-temporally align sequences of deformable objects is that the reliance on manual annotations can be minimized.

The major challenge of performing joint spatio-temporal alignment of sequences that display texture-varying deformable objects is the high dimensionality of the texture space. Hence, we need to device component analysis methodologies that can extract a small number of components suitable for both spatial and temporal alignment. Then, spatial non-rigid, as well as temporal, alignment can be conducted using the low-dimensional space. In this paper, motivated by the recent success on combining component analysis with (i) spatial non-rigid deformations by means of a statistical shape model for deformable alignment of image sets [21, 2, 29, 9, 33], and (ii) temporal deformations by means of Dynamic Time Warping (DTW) [34, 30, 17], we propose, the first, to the best of our knowledge, component analysis methodology which can perform joint spatio-temporal alignment of two sequences.

The proposed methodology is radically different compared to recent methods that perform joint component analysis and spatial alignment [21, 2, 29, 9, 33]. Specifically, our technique is totally different than [21, 9, 33], which are based on pre-trained models of appearance and require annotations of hundreds of images in order to achieve good generalization properties. The most closely related methods are the unsupervised method of [29] that performs component analysis for unsupervised non-rigid spatial alignment and the method of [34] that performs temporal alignment. The component analysis methodology used in [34] for joint dimensionality reduction and temporal alignment is based on Canonical Correlation Analysis (CCA). CCA does not use any temporal model or regularization, and most importantly, due to generalized orthogonality constraints, does not provide good reconstruction of the sequences. Hence, it is not ideal for spatial alignment (in Sec. 2.5.1 we thoroughly discuss the relationship of the proposed methodology with CCA). Similarly, the recently proposed temporally regularized Principal Component Analysis (PCA) (so called Autoregressive Component Analysis (ARCA)) [29] is tailored only to preserve the reconstruction of the sequence's images, without discovering common low-dimensional features that can be used for temporal alignment.

2. Method

In this section, we start by reviewing the spatial (Sec. 2.1) and temporal (Sec. 2.2) alignment methods of image sequences that are closely related to the proposed technique. Then, we present our method for describing a spatio-temporal phenomenon (Sec. 2.3). Finally, we discuss its convergence (Sec. 2.4), its relationship with existing CCA techniques and give a probabilistic interpretation (Sec. 2.5).

2.1. Unsupervised Deformable Spatial Alignment of Image Sequences

Recently, the line of research of joint component analysis and spatial alignment has received attention [21, 2, 29, 9, 33]. Some of the methods require a known set of bases that is build from a set of already aligned objects [21, 9, 33]. In this paper, we are interested in the unsupervised alignment of image sequences. The most recently proposed method for that task is [29]. In that work, it is assumed that only a statistical model of the facial shape is given. Let us express a shape instance that comprises of a set of S landmarks as $\mathbf{s} = [x_1, y_1, \ldots, x_S, y_S]^T$, where (x_i, y_i) are the coordinates that correspond to the i-th landmark. A statistical shape model can be easily learned by performing PCA on a set of training shapes in order to acquire a set of bases $\mathbf{U}_S$ and the mean shape $\bar{\mathbf{s}}$. A new shape instance can be approximately parametrised using the learned model, as $\mathbf{s}_t \approx \bar{\mathbf{s}} + \mathbf{U}_S\mathbf{p}$, where $\mathbf{p}$ is the set of parameters. Rigid transformations can be incorporated in the bases $\mathbf{U}_S$ [16]. Given an image and a vector of parameters $\mathbf{p}$ that describes a shape instance in the image, then the texture of the image can be warped into a predefined reference frame. In this paper, we denote the warped image as $\mathbf{x}(\mathbf{p})$. The warp can be formulated in two ways: (i) as a non-linear function, such as Piece-Wise Affine (PWA), in order to sample the whole image, and (ii) as a simple translational model [25] that samples only the local texture around landmarks.

Given a set of N images stacked as the columns of a matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, the method proposed in [29] (so called ARCA) learns a temporally regularized decomposition of $\mathbf{X}$ and, at the same time, estimates the shapes of the faces included in the images by extracting a set of parameters $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_N]$. The optimization problem is

$$\mathbf{P}^o, \mathbf{U}^o, \mathbf{V}^o = \underset{\mathbf{P},\mathbf{U},\mathbf{V}}{\operatorname{argmin}} ||\mathbf{X}(\mathbf{P}) - \mathbf{U}\mathbf{V}||_F^2 + \lambda \operatorname{tr}[\mathbf{V}\mathbf{L}\mathbf{V}^T] \quad (1)$$

where $\mathbf{X}(\mathbf{P}) = [\mathbf{x}_1(\mathbf{p}_1), \ldots, \mathbf{x}_N(\mathbf{p}_N)]$, and $||.||_F^2$ and $\operatorname{tr}[.]$ denote the squared Frobenius norm of matrices and the trace matrix operator, respectively. Finally

$$\mathbf{L} = \begin{pmatrix} 1 & -\phi & & & \\ -\phi & 1+\phi^2 & -\phi & & \\ & \ddots & \ddots & \ddots & \\ & & -\phi & 1+\phi^2 & -\phi \\ & & & -\phi & 1 \end{pmatrix} \quad (2)$$

is an appropriate Laplacian matrix that incorporates first order Markov dependencies between data. The authors in [29]

follow an alternating minimization procedure and show that the above optimization problem, not only can provide a non-rigid alignment of the images, but the weights $\mathbf{V}$ contain smooth information that can be used to perform unsupervised analysis of facial behaviour (i.e., segment facial expressions with regards to several temporal segments). Furthermore, they explore the relationship between the above model and Slow Feature Analysis (SFA) [27].

2.2. Temporal Alignment of Image Sequences

DTW [15] is a popular algorithm for the temporal alignment of two sequences that have different lengths. In particular, given two sequences stored as the columns of two matrices $\mathbf{X}_1 \in \mathbb{R}^{F\times N_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{F\times N_2}$, where N_1 and N_2 are the respective number of frames, DTW finds two binary warping matrices $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$ so that the least squares error between the warped sequences is minimised. This is expressed as

$$\begin{aligned}\mathbf{\Delta}_{1,2}^o = &\underset{\mathbf{\Delta}_{1,2}}{\operatorname{argmin}} \|\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{X}_2\mathbf{\Delta}_2\|_F^2 \\ &\text{s.t. } \mathbf{\Delta}_1 \in \{0,1\}^{T_1\times T},\ \mathbf{\Delta}_2 \in \{0,1\}^{T_2\times T}\end{aligned} \tag{3}$$

where T is the length of the common aligned path. DTW is able to find the optimal alignment path by using dynamic programming [4] despite the fact that the number of possible alignments is exponential with respect to T_1 and T_2.

However, DTW has some important limitations. Firstly, it is largely affected by the dimensionality of the data and, secondly, it is not able to align signals of different dimensions. In order to accommodate for the above, as well as for differences regarding the nature, style and subject variability of the signals, Canonical Time Warping (CTW) was proposed in [34]. CTW combines DTW with CCA, in order to add a principled feature selection and dimensionality reduction mechanism within DTW. In particular, by taking advantage of the similarities between the least squares functional form of CCA [10] and Eq. 3, CTW simultaneously discovers two linear operators ($\mathbf{U}_1$, $\mathbf{U}_2$) and applies DTW on the low dimensional embedding of $\mathbf{U}_1^T\mathbf{X}_1$ and $\mathbf{U}_2^T\mathbf{X}_2$ by solving the following optimization problem

$$\begin{aligned}\mathbf{\Delta}_{1,2}^o, \mathbf{U}_{1,2}^o = &\underset{\mathbf{\Delta}_{1,2},\mathbf{U}_{1,2}}{\operatorname{argmin}} \|\mathbf{U}_1^T\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{U}_2^T\mathbf{X}_2\mathbf{\Delta}_2\|_F^2 \\ &\text{s.t. } \mathbf{\Delta}_1 \in \{0,1\}^{T_1\times T},\ \mathbf{\Delta}_2 \in \{0,1\}^{T_2\times T} \\ &\mathbf{U}_1^T\mathbf{X}_1\mathbf{D}_1{\mathbf{X}_1}^T\mathbf{U}_1 = \mathbf{I},\ \mathbf{U}_2^T\mathbf{X}_2\mathbf{D}_2{\mathbf{X}_2}^T\mathbf{U}_2 = \mathbf{I}\end{aligned} \tag{4}$$

where $\mathbf{D}_1 = \mathbf{\Delta}_1\mathbf{\Delta}_1^T$ and $\mathbf{D}_2 = \mathbf{\Delta}_2\mathbf{\Delta}_2^T$. An alternating optimization approach was used in order to solve the above problem. One of the drawbacks of CTW is that it does not take into account the dynamic information of the signals. Furthermore, even though CTW can theoretically handle high dimensional spaces, in [34] it has only been tested on alignment problems that deal with sparse sets of landmarks. According to our experiments, for the task of aligning facial behaviour using image pixel information, CTW can perform well only if a dimensionality reduction step has been applied on each video using PCA.

2.3. A Correlated Component Analysis for Describing a Spatio-Temporal Phenomenon

In this section, we build on the component analysis model of ARCA [29] in order to describe a temporal phenomenon which is common in two sequences that are both spatially and temporally aligned (e.g. two video sequences depicting the same expression or Facial Action Unit (AU) [13]). Then, we assume that the sequences' frames are neither spatially nor temporally aligned and we propose an optimization problem that jointly decomposes the image sequences into maximally correlated subspaces and performs spatial and temporal alignment.

Let us denote two image sequences as the stacked matrices $\mathbf{X}_1 = [\mathbf{x}_1^1, \ldots, \mathbf{x}_N^1]$ and $\mathbf{X}_2 = [\mathbf{x}_1^2, \ldots, \mathbf{x}_N^2]$. We assume that both sequences are explained by a linear generative model. That is, we want to decompose the two sequences into two maximally correlated subspaces $\mathbf{V}_1$ and $\mathbf{V}_2$ using the orthonormal bases $\mathbf{U}_1$ and $\mathbf{U}_2$, as

$$\begin{aligned}\mathbf{U}_{1,2}^o, \mathbf{V}_{1,2}^o = &\underset{\mathbf{U}_{1,2},\mathbf{V}_{1,2}}{\operatorname{argmin}} ||\mathbf{X}_1 - \mathbf{U}_1\mathbf{V}_1||_F^2 + ||\mathbf{X}_2 - \mathbf{U}_2\mathbf{V}_2||_F^2 \\ &+ \lambda\text{tr}[\mathbf{V}_1\mathbf{L}\mathbf{V}_1^T] + \lambda\text{tr}[\mathbf{V}_2\mathbf{L}\mathbf{V}_2^T] + ||\mathbf{V}_1 - \mathbf{V}_2||_F^2 \\ &\text{s.t. } \mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I},\ \mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I}\end{aligned} \tag{5}$$

In Sec. 2.5.1 we show how the component analysis is linked to CCA and explore the main modelling differences.

Assuming that the sequences $\mathbf{X}_1 \in \mathbb{R}^{F\times N_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{F\times N_2}$ are neither temporally aligned, hence they do not have the same length, nor spatially aligned, we propose the following optimization problem

$$\begin{aligned}&\mathbf{P}_{1,2}^o, \mathbf{\Delta}_{1,2}^o, \mathbf{U}_{1,2}^o, \mathbf{V}_{1,2}^o = \\ &= \underset{\mathbf{P}_{1,2},\mathbf{\Delta}_{1,2},\mathbf{U}_{1,2},\mathbf{V}_{1,2}}{\operatorname{argmin}} ||(\mathbf{X}_1(\mathbf{P}_1) - \mathbf{U}_1\mathbf{V}_1)\mathbf{\Delta}_1||_F^2 + \\ &+ ||(\mathbf{X}_2(\mathbf{P}_2) - \mathbf{U}_2\mathbf{V}_2)\mathbf{\Delta}_2||_F^2 + \lambda\text{tr}[\mathbf{V}_1\mathbf{L}_1\mathbf{V}_1^T] + \\ &+ \lambda\text{tr}[\mathbf{V}_2\mathbf{L}_2\mathbf{V}_2^T] + ||\mathbf{V}_1\mathbf{\Delta}_1 - \mathbf{V}_2\mathbf{\Delta}_2||_F^2 \\ &\text{s.t. } \mathbf{\Delta}_1 \in \{0,1\}^{N_1\times N},\ \mathbf{\Delta}_2 \in \{0,1\}^{N_2\times N} \\ &\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I},\ \mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I}\end{aligned} \tag{6}$$

where $\mathbf{L}_1 \in \mathbb{R}^{N_1\times N_1}$ and $\mathbf{L}_2 \in \mathbb{R}^{N_2\times N_2}$ are Laplacian matrices and $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$ are binary warping matrices. The above optimization problem forms the bases of our framework and enables us to perform joint spatio-temporal alignment of the sequences into a common frame defined as the mean shape $\bar{\mathbf{s}}$. In Section 2.5 we discuss the relationship between the above model and CCA/CTW. The advantages of the proposed model over CTW is (a) the proposed model

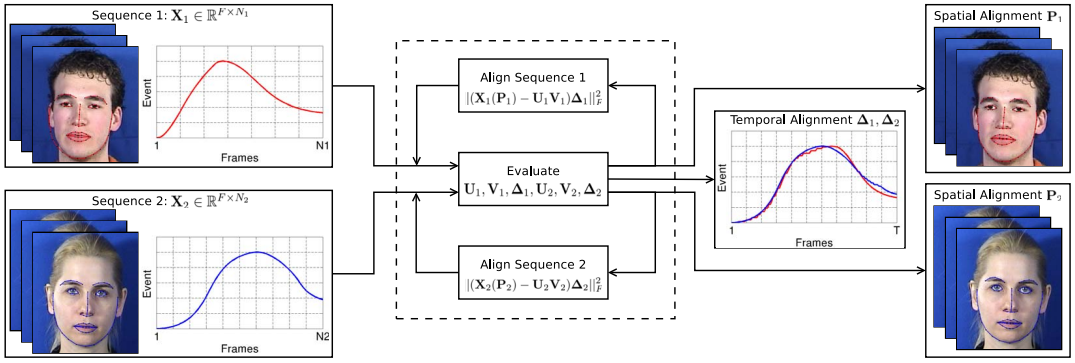


Figure 1: Method overview. Given two video sequences, the proposed method performs joint deformable spatio-temporal alignment using an iterative procedure that gradually improves the result. The initialization is acquired by applying ARCA [29] on both sequences.

incorporates temporal regularisation contraints, (b) we can perform jointly temporal and spatial alignment and (c) we can easily incorporate terms that account for gross corruptions/error [18].

The sequences that consist of the warped frames' vectors are given by

$$\mathbf{X}_i(\mathbf{P}_i) = \left[\mathbf{x}_1^i(\mathbf{p}_1^i), \ldots, \mathbf{x}_{N_i}^i(\mathbf{p}_{N_i}^i)\right], \; i = 1, 2 \tag{7}$$

where $\mathbf{P}_i = [\mathbf{p}_1^i, \ldots, \mathbf{p}_{N_i}^i]$ is the matrix of the shape parameters of each frame and i denotes the sequence index. As shown in overview of Fig. 1, the above optimization problem is iteratively solved in an alternating manner. The first step is to estimate matrices $\mathbf{U}_{1,2}$ and $\mathbf{V}_{1,2}$ based on the current estimate of the shape parameters $\mathbf{P}_{1,2}$ and then apply DTW on $\mathbf{V}_{1,2}$ in order to find $\mathbf{\Delta}_{1,2}$. The second step is to compute the parameters of the spatial alignment $\mathbf{P}_{1,2}$ given the current estimation of $\mathbf{U}_{1,2}$, $\mathbf{V}_{1,2}$ and $\mathbf{\Delta}_{1,2}$. The initial shapes are estimated by applying ARCA on both sequences. Therefore, the optimization of Eq. 6 is solved in the following two steps:

2.3.1 Fix $\mathbf{P}_{1,2}$ and minimize with respect to $\mathbf{U}_{1,2}, \mathbf{V}_{1,2}$ and $\mathbf{\Delta}_{1,2}$

In this step of the proposed method, we aim to update $\mathbf{U}_{1,2}$ and $\mathbf{V}_{1,2}$, assuming that we have a current estimate of the shape parameters' matrices $\mathbf{P}_{1,2}$, hence of the data matrices $\mathbf{X}_{1,2}(\mathbf{P}_{1,2})$. Those updates are estimated by using an alternating optimization framework. Specifically, we first fix $\mathbf{V}_{1,2}$ and compute $\mathbf{U}_{1,2}$ and then we find $\mathbf{V}_{1,2}$ by fixing $\mathbf{U}_{1,2}$. The warping matrices $\mathbf{\Delta}_{1,2}$ are updated at the beginning of each such iteration.

Update $\mathbf{\Delta}_{1,2}$ In the first iteration, we assume that we have the initial $\mathbf{V}_{1,2}$ obtained by applying the ARCA algorithm on each sequence $\mathbf{X}_{1,2}(\mathbf{P}_{1,2})$. Thus, the warping matrices $\mathbf{\Delta}_{1,2}$ are estimated by applying DTW on these initial $\mathbf{V}_{1,2}$. In every subsequent iteration, $\mathbf{\Delta}_{1,2}$ are estimated by applying DTW on the updated $\mathbf{V}_{1,2}$, thus $(\mathbf{\Delta}_1, \mathbf{\Delta}_2) = \text{DTW}(\mathbf{V}_1, \mathbf{V}_2)$.

Update $\mathbf{U}_{1,2}$ Given the current estimate of $\mathbf{V}_{1,2}$, the optimization problem with regards to $\mathbf{U}_{1,2}$ is given by

$$\begin{aligned} f(\mathbf{V}_{1,2}) =& \|(\mathbf{X}_1(\mathbf{P}_1) - \mathbf{U}_1\mathbf{V}_1)\mathbf{\Delta}_1\|_F^2 + \|(\mathbf{X}_2(\mathbf{P}_2) - \mathbf{U}_2\mathbf{V}_2)\mathbf{\Delta}_2\|_F^2 \\ &\text{s.t. } \mathbf{\Delta}_1 \in \{0,1\}^{N_1 \times N}, \mathbf{\Delta}_2 \in \{0,1\}^{N_2 \times N} \\ &\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}, \; \mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I} \end{aligned} \tag{8}$$

The updates from the above optimization problem are derived by the Skinny Singular Value Decomposition (SSVD) [37] of $\mathbf{X}_i(\mathbf{P}_i)\mathbf{D}_i\mathbf{V}_i^T$. That is, given the SVD $\mathbf{X}_i(\mathbf{P}_i)\mathbf{D}_i\mathbf{V}_i^T = \mathbf{R}_i\mathbf{S}_i\mathbf{M}_i^T$, then

$$\mathbf{U}_i = \mathbf{R}_i\mathbf{M}_i^T, \; i = 1, 2 \tag{9}$$

where, for convenience, we set $\mathbf{D}_i = \mathbf{\Delta}_i\mathbf{\Delta}_i^T, \; i = 1, 2$

Update $\mathbf{V}_{1,2}$ Given $\mathbf{U}_{1,2}$, the optimization problem with regards to $\mathbf{V}_{1,2}$ is formulated as

$$\begin{aligned} f(\mathbf{U}_{1,2}) =& \|(\mathbf{X}_1(\mathbf{P}_1) - \mathbf{U}_1\mathbf{V}_1)\mathbf{\Delta}_1\|_F^2 + \lambda \text{tr}[\mathbf{V}_1\mathbf{L}_1\mathbf{V}_1^T] \\ &+ \|(\mathbf{X}_2(\mathbf{P}_2) - \mathbf{U}_2\mathbf{V}_2)\mathbf{\Delta}_2\|_F^2 + \lambda \text{tr}[\mathbf{V}_2\mathbf{L}_2\mathbf{V}_2^T] \\ &+ \|\mathbf{V}_1\mathbf{\Delta}_1 - \mathbf{V}_2\mathbf{\Delta}_2\|_F^2 \\ &\text{s.t. } \mathbf{\Delta}_1 \in \{0,1\}^{N_1 \times N}, \; \mathbf{\Delta}_2 \in \{0,1\}^{N_2 \times N} \\ &\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}, \; \mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I} \end{aligned} \tag{10}$$

By evaluating the partial derivatives with respect to $\mathbf{V}_i, \; \forall i = \{1, 2\}$ and equalize them with zero, we derive

$$\mathbf{V}_i = (\mathbf{U}_i^T\mathbf{X}_i\mathbf{D}_i + \mathbf{C}_i)(2\mathbf{D}_i + \lambda\mathbf{L}_i)^{-1}, \; i = 1, 2 \tag{11}$$

where $\mathbf{C}_1 = \mathbf{V}_2\mathbf{\Delta}_2\mathbf{\Delta}_1^T$ and $\mathbf{C}_2 = \mathbf{V}_1\mathbf{\Delta}_1\mathbf{\Delta}_2^T$.

2.3.2 Fix $\mathbf{U}_{1,2}, \mathbf{V}_{1,2}, \mathbf{\Delta}_{1,2}$ and minimize with respect to $\mathbf{P}_{1,2}$

The aim of this step is to estimate the shape parameters' matrices $\mathbf{P}_i, \; i = 1, 2$ for each sequence, given the current

estimate of the bases $\mathbf{U}_{1,2}$ and the features $\mathbf{V}_{1,2}$. This is performed for each sequence independently and it can be expressed as the following optimization problem

$$\begin{aligned}\mathbf{P}_i^o &= \underset{\mathbf{P}_i}{\operatorname{argmin}} \|\mathbf{X}_i(\mathbf{P}_i) - \mathbf{U}_i\mathbf{V}_i\|_F^2 = \\ &= \underset{\{\mathbf{p}_j^i\},\, j=1,\ldots,N_i}{\operatorname{argmin}} \sum_{j=1}^{N_i} \|\mathbf{x}_j^i(\mathbf{p}_j^i) - \mathbf{U}_i\mathbf{v}_j^i\|_2^2,\ i=1,2\end{aligned} \tag{12}$$

where $\mathbf{v}_j^i,\ \forall j = 1,\ldots,N_i,\ \forall i = 1,2$ denotes the j-th column of the matrix $\mathbf{V}_i$ that corresponds to each sequence. In other words, for each sequence ($i = 1,2$), we aim to minimize the Frobenius norm between the warped frames $\mathbf{X}_i(\mathbf{P}_i)$ and the templates $\mathbf{U}_i\mathbf{V}_i$. The solution is obtained by employing the Inverse Compositional (IC) Image Alignment algorithm [3]. Note that the IC alignment is performed separately for each frame of each sequence. In brief, the solution can be derived by introducing an incremental warp term ($\Delta\mathbf{p}_j^i$) on the part of the template of Eq. 12. Then, by linearizing (first order Taylor expansion) around zero ($\Delta\mathbf{p}_j^i = \mathbf{0}$), the incremental warp is given by

$$\Delta\mathbf{p}_j^i = \mathbf{H}^{-1}\mathbf{J}^T|_{\mathbf{p}=\mathbf{0}} \left[\mathbf{x}_j^i(\mathbf{p}_j^i) - \mathbf{U}_i\mathbf{v}_j^i\right], \quad j = 1,\ldots,N_i,\ i = 1,2$$

where $\mathbf{H} = \mathbf{J}^T|_{\mathbf{p}=\mathbf{0}}\mathbf{J}|_{\mathbf{p}=\mathbf{0}}$ is the Gauss-Newton approximation of the Hessian matrix and $\mathbf{J}^T|_{\mathbf{p}=\mathbf{0}}$ is the Jacobian of each template $\mathbf{U}_i\mathbf{v}_j^i$. The biggest advantage of the IC algorithm is that the Jacobian and the inverse of the Hessian matrix are constant and can be precomputed once, because the linearization of the solution is taken on the part of the template.

2.4. Empirical Convergence

Herein, we empirically investigate the convergence of the proposed optimization problem on MMI and UNS databases. Figure 2 shows the values of the cost function of Eq. 6, averaged over all the videos. The results show that the proposed methodology converges monotonically and 4-5 iterations are adequate to achieve good performance.

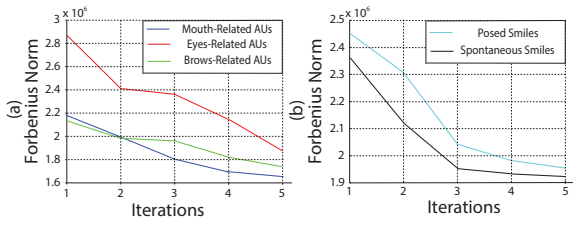


Figure 2: Cost function error with respect to the iterations averaged over all (a) MMI and (b) UNS videos.

2.5. Theoretical Interpretation

2.5.1 Relationship to Canonical Component Analysis

In this section, we analyze the relationship between the proposed model of Sec. 2.3 and other methodologies that produce subspaces of correlated features. Naturally, this comparison is mostly targeted on the close related CCA. Let us formulate the optimization problem of Eq. 5 without the temporal regularization terms, as

$$\begin{aligned}\mathbf{U}_{1,2}^o, \mathbf{V}_{1,2}^o = \underset{\mathbf{U}_1,\mathbf{U}_2,\mathbf{V}_1,\mathbf{V}_2}{\operatorname{argmin}} &||\mathbf{X}_1 - \mathbf{U}_1\mathbf{V}_1||_F^2 + \\ &+ ||\mathbf{X}_2 - \mathbf{U}_2\mathbf{V}_2||_F^2 + ||\mathbf{V}_1 - \mathbf{V}_2||_F^2 \\ &\text{s.t. } \mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I},\ \mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I}\end{aligned} \tag{13}$$

By assuming that the weights matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ are formed by projecting the sequences onto the respective orthonormal bases as $\mathbf{V}_1 = \mathbf{U}_1^T\mathbf{X}_1$ and $\mathbf{V}_2 = \mathbf{U}_2^T\mathbf{X}_2$, and then substituting back to Eq. 13, we end up with

$$\begin{aligned}\mathbf{U}_{1,2}^o = &\underset{\mathbf{U}_1,\mathbf{U}_2}{\operatorname{argmax}} \operatorname{tr}\left[\begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}^T \begin{pmatrix}\mathbf{0} & \mathbf{X}_1\mathbf{X}_2^T\\ \mathbf{X}_2\mathbf{X}_1^T & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}\right] \\ &\text{s.t. } \begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}^T \begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix} = \mathbf{I}\end{aligned} \tag{14}$$

which is a special case of CCA with orthogonal instead of generalized orthogonal constraints[1]. The derivation of the above problem is shown in the supplementary material and its solution is given by performing eigen-analysis.

Motivated by Eq. 14, it can be shown that the proposed component analysis formulation of Eq. 6 is a case of orthogonal CCA with temporal regularized terms. Specifically, by assuming that $\mathbf{V}_1 = \mathbf{U}_1^T\mathbf{X}_1$ and $\mathbf{V}_2 = \mathbf{U}_2^T\mathbf{X}_2$, the optimization problem of Eq. 6 can be reformulated as

$$\begin{aligned}&\mathbf{U}_1^o, \mathbf{U}_2^o = \\ &\underset{\mathbf{U}_1,\mathbf{U}_2}{\operatorname{argmax}} \operatorname{tr}\left[\begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}^T \begin{pmatrix}-\mathbf{X}_1\mathbf{L}\mathbf{X}_1^T & \mathbf{X}_1\mathbf{X}_2^T\\ \mathbf{X}_2\mathbf{X}_1^T & -\mathbf{X}_2\mathbf{L}\mathbf{X}_2^T\end{pmatrix}\begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}\right] \\ &\text{s.t. } \begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}^T \begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix} = \mathbf{I}\end{aligned} \tag{15}$$

which again can be solved by performing eigen-analysis. The above problem is a kind of temporally regularized orthogonal CCA.Temporal regularisation is probably the reason that the proposed approach outperforms CTW (which does not employ any temporal regularisation).

Even though Laplacian regularization of component analysis techniques has recently been significantly studied [7], Laplacian regularization for CCA models has not received much attention [5]. To the best of our knowledge, this is the first component analysis methodology which can

[1] CCA has as constraints $\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{U}_1 = \mathbf{I}$ and $\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{U}_2 = \mathbf{I}$.

lead to a CCA with temporal regularization terms[2]. We believe that the proposed component analysis method is superior to the CCA model for both spatial and temporal alignment, since (a) the bases are orthogonal and hence can be used to build better statistical models for spatial alignment [16] and (b) we have applied temporal regularization terms which produce smoother latent spaces $\mathbf{V}_1$ and $\mathbf{V}_2$ which are better for temporal alignment. Finally, note that the reason why we solve the proposed decomposition using the least-squares approach and not eigen-analysis is numerical stability [10].

2.5.2 Probabilistic Interpretation

The proposed optimization problem also provides the maximum-likelihood solution of a shared space generative autoregressive model. That is, we assume we have two linear models that describe the generation of observations in the two sequences

$$\begin{aligned} \mathbf{x}_i^1 &= \mathbf{U}_1\mathbf{v}_i^1 + \mathbf{e}_i^1,\ \mathbf{e}_i^1 \sim \mathcal{N}(\mathbf{0}, \sigma_1\mathbf{I}),\ i = 1, \ldots, N_1 \\ \mathbf{x}_i^2 &= \mathbf{U}_2\mathbf{v}_i^2 + \mathbf{e}_i^2,\ \mathbf{e}_i^2 \sim \mathcal{N}(\mathbf{0}, \sigma_2\mathbf{I}),\ i = 1, \ldots, N_2 \end{aligned} \tag{16}$$

Let us also make the assumption that $\mathbf{V}_1 = [\mathbf{v}_1^1, \ldots, \mathbf{v}_{N_1}^1]$ forms an autoregressive sequence. That is, $\mathbf{V}_1 \sim \frac{|\mathbf{L}^N|}{\sqrt{(2\pi)^{kN}}} \exp\{-\frac{1}{2}\text{tr}[\mathbf{V}_1\mathbf{L}\mathbf{V}_1^T]\}$ with $\mathbf{L}$ being the Laplacian and $\mathbf{V}_2$ is the same as $\mathbf{V}_1$ up to a Gaussian noise, i.e. $\mathbf{v}_i^1 = \mathbf{v}_i^2 + \mathbf{e}_i$ with $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$. It is straightforward to show that maximizing the joint log likelihood of the above probabilistic model with regards to $\mathbf{U}_1, \mathbf{U}_2, \mathbf{V}_1$ and $\mathbf{V}_2$ is equivalent to optimizing the cost function in Eq. 13.

It is worthwhile to compare the proposed with the Dynamic Probabilistic CCA (DPCCA) method proposed in [17]. The method in [17] models shared and individual spaces in a probabilistic manner, i.e. by incorporating priors over these spaces and marginalising them out. Time series alignment is performed by applying DTW on the expectations of the shared space over the individual posteriors. Using the model in [17] to perform joint spatial alignment is not trivial, that is why temporal alignment is performed on facial shape only.

3. Experiments

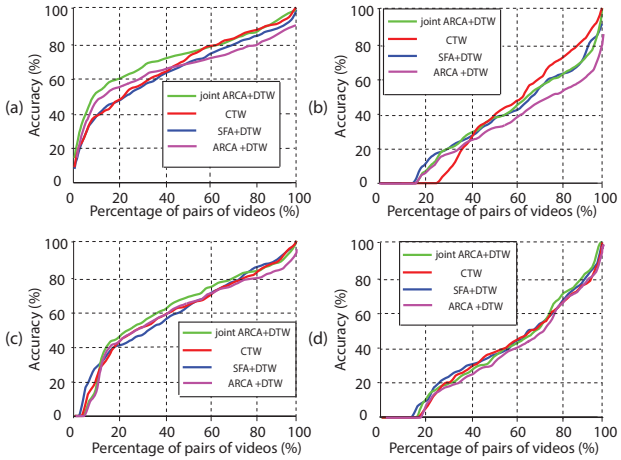
In order to demonstrate the effectiveness of the proposed framework, we conduct experiments on two datasets: MMI [20, 26] which consists of videos with posed AUs and UvA-Nemo Smile (UNS) [11] which contains videos with posed and spontaneous smiles. The MMI database contains more than 400 videos, in which a subject performs one or more AUs that are annotated with respect to the following temporal segments: (1) *neutral* when there is no facial motion, (2) *onset* when the facial motion starts, (3) *apex*, when the muscles reach the peak intensity, and (4) *offset* when the muscles begin to relax. The large-scale UNS database consists of more than 1240 videos (597 spontaneous and 643 posed) with 400 subjects. Since this database does not provide any annotations of temporal segments, we manually annotated 50 videos displaying spontaneous smiles and 50 videos displaying posed smiles using the same temporal segments as in the case of MMI.

3.1. Temporal Alignment Results

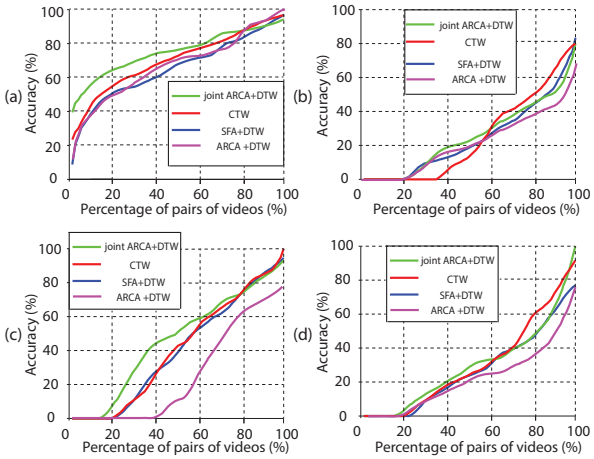
In this section, we provide experimental results for the temporal alignment of pairs of videos from both the MMI and UNS databases. The pairs are selected so that the same AU is activated. The aim of those experiments is (a) to evaluate the performance of the proposed framework compared to various commonly used temporal alignment methods, and (b) to show that by treating the problems of spatial and temporal alignment jointly instead of independently we achieve better results. We compare the proposed unsupervised framework, labelled as *joint ARCA+DTW*, with **(a)** *CTW*, **(b)** *SFA+DTW*, and **(c)** *ARCA+DTW* in which the problems of temporal and spatial alignment are solved independently. For the joint ARCA+DTW, we set the parameter λ that regulates the contribution of the smoothness constraints equal to 150 for both sequences. Furthermore, the dimensionality of the latent space for all the examined methods is set to 25, which was the dimensionality that lead to the best performance in a validation set. The matrices were initialised by applying first ARCA on both sequences. The shape parameters were initialised with zeros and the mean shape was placed in the bounding box returned by Viola-Jones face detector [31]. Finally, the proposed method is applied for 5 global iterations. We would like to note that we have run *ARCA+DTW* one and several iterations but because there is no joint subspace learned between two videos we have not observed any improvement.

The temporal alignment accuracy is evaluated by employing the metric used in recent works [17]. Specifically, let us assume that we have 2 video sequences with the corresponding features ($\mathbf{V}_i,\ i = 1, 2$) and AU annotations ($\mathbf{A}_i,\ i = 1, 2$). Additionally, assume that we have recovered the alignment binary matrices $\mathbf{\Delta}_i,\ i = 1, 2$ for each video. By applying these matrices on the AU annotations (i.e., $\mathbf{A}_1\mathbf{\Delta}_1$ and $\mathbf{A}_2\mathbf{\Delta}_2$) we can find the temporal phase of the AU that each aligned frame of each video corresponds to. Therefore, for a given temporal phase (e.g., neutral), we have a set of frame indices which are assigned to the specific temporal phase in each video, i.e. $\mathcal{N}_1^p$ and $\mathcal{N}_2^p$ respectively.

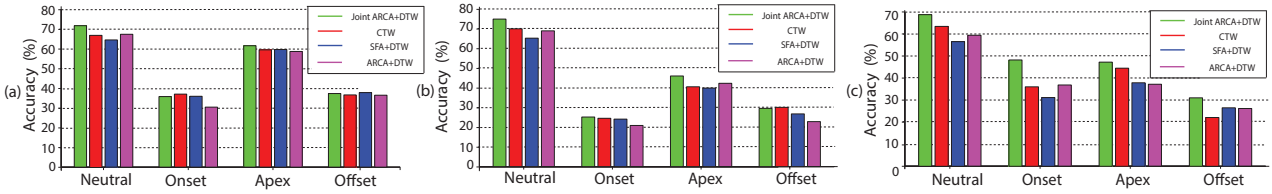
[2]Our component analysis is not to be confused with the so-called Dynamic CCA model proposed in [17], where special probabilistic Linear Dynamical Systems (LDS) are proposed with shared and common spaces. The proposed model is deterministic. It is also radically different to the so-called semi-supervised Laplacian CCA method of [5], where a semi-supervised Linear Discriminant Analysis (LDA) is proposed.



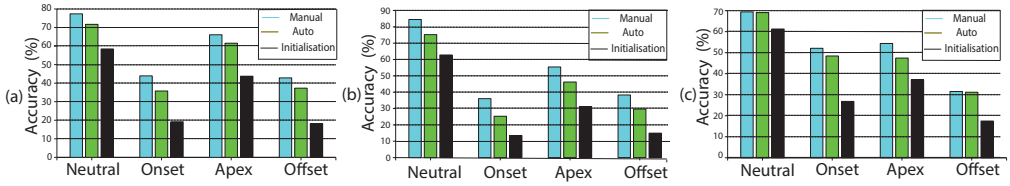
(**i**) Percentage of video pairs that achieve an accuracy less or equal than the respective value for mouth-related AUs. The subfigures correspond to the temporal phases as: *(a)* neutral, *(b)* onset, *(c)* apex, *(d)* offset.



(**ii**) Percentage of video pairs that achieve an accuracy less or equal than the respective value for eyes-related AUs. The subfigures correspond to the temporal phases as: *(a)* neutral, *(b)* onset, *(c)* apex, *(d)* offset.



(**iii**) Average accuracy over all the video pairs with respect to the temporal phase for *(a)* mouth-related AUs, *(b)* eyes-related AUs *(c)* brows-related AUs.



(**iv**) Average accuracy of the proposed method for different spatial alignment scenarios for *(a)* mouth-related AUs, *(b)* eyes-related AUs, *(c)* brows-related AUs. *Auto*: Proposed joint unsupervised spatial alignment. *Manual*: Using the manually annotated landmarks. *Initialisation*: Random initialisation.

Figure 3: Temporal alignment results on MMI database.

The accuracy is then estimated as $\frac{|\mathcal{N}_1^p \cap \mathcal{N}_2^p|}{|\mathcal{N}_1^p \cup \mathcal{N}_2^p|}$, which essentially corresponds to the ratio of correctly aligned frames to the total duration of the temporal phase p across the aligned videos.

3.1.1 MMI database

In this section, we report the performance on MMI database. The experiments are conducted on 480 pairs of videos that depict the same AU. The results are split in three categories, based on the region of the face that is activated by the performed AU, i.e. *mouth*, *eyes* and *brows*. For each facial region, the results are further separated per temporal segment. The AUs that correspond to each facial region are:

• **Mouth**: Upper Lip Raiser, Nasolabial Deepener, Lip Corner Puller, Cheek Puffer, Dimpler, Lip Corner Depressor, Lower Lip Depressor, Chin Raiser, Lip Puckerer, Lip Stretcher, Lip Funneler, Lip Tightener, Lip Pressor, Lips Part, Jaw Drop, Mouth Stretch, Lip Suck

• **Eyes**: Upper Lid Raiser, Cheek Raiser, Lid Tightener, Nose Wrinkler, Eyes Closed, Blink, Wink, Eyes Turn Left and Eyes Turn Right

• **Brows**: Inner Brow Raiser, Outer Brow Raiser and Brow Lowerer

Figure 3 summarizes the temporal alignment of three experiments on the MMI database. Specifically, Figures 3**i** and 3**ii** show the percentage of video pairs that achieved an accuracy less or equal than the corresponding value for mouth-related and eyes-related AUs, respectively. In other words, these Cumulative Accuracy Distributions (CAD) show the percentage of video pairs that achieved at most a specific accuracy percentage. The plots for each facial region are also separated with respect to the temporal segment in question. The results indicate that, for both mouth

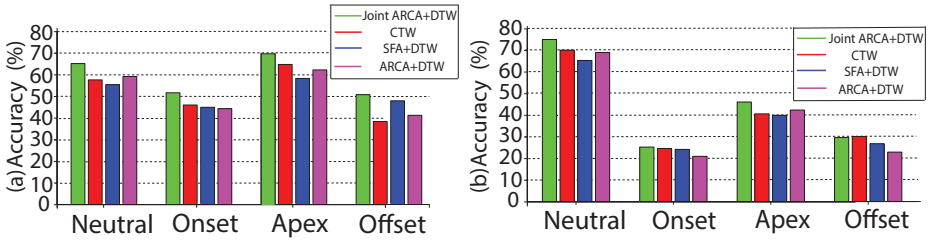


Figure 4: Average accuracy over all the video pairs with respect to the temporal phase for *(a)* spontaneous smiles, *(b)* posed smiles.

and eyes related AUs, our method outperforms the rest of techniques for the neutral and apex phases, and has a comparable performance for onset and offset.

This is better illustrated in Fig. 3**iii** which reports the average accuracy over all the video pairs for each temporal phase separately. The results for the brows-related AUs are also included in this figure, which indicate that the proposed method significantly outperforms the other techniques for all the temporal phases. Due to limited space, the CAD curves for the brows-related AUs for each temporal phase is omitted and can be found in the supplementary material. Moreover, note that our methodology outperforms ARCA+DTW for all facial regions and temporal phases. This is an important result which indicates that treating the spatial and temporal alignment as a joint problem is more advantageous than solving them independently.

Regarding the alignment of mouth-related AUs, it is worth mentioning that a similar experiment with the one provided in this section (Fig. 3**iii** (a)) was conducted in [17] (section 7.3), which reports the average accuracy over 50 video pairs performing AU12 in MMI database. Specifically for this task, we obtained 71% accuracy over DPCTW which obtained 55% for the neutral phase in the features. Subsequently, we achieved 38% accuracy compared to 33% for the onset phase, 61% over 60% for the apex phase and 39% compared to 37% for the offset phase. We have to note that our algorithm is completely automatic in terms of both spatial and temporal alignment (requiring only a face detector) and uses raw pixel intensities. On the other hand the method in [17] used, manually corrected, tracked landmarks.

Figure 3**iv** reports the results of a second experiment that aims to assess the effect of spatial alignment in the temporal alignment procedure. Specifically, we apply the proposed technique with different spatial alignment approaches, that is (a) the proposed unsupervised spatial alignment, (b) using the manually annotated landmarks, (c) adding random noise to the manually annotated landmarks. The results indicate that in most cases, the proposed method with automatic spatial alignment greatly outperforms the case of random initialisation and has comparable performance with the case of perfectly aligned images.

3.1.2 UNS database

In this section, we provide temporal alignment results on the UNS database, which contains, not only posed, but also spontaneous smiles which are more complex due to their dynamics [32]. We conduct the experiments on 188 pairs of videos with posed smiles and 122 pairs with spontaneous smiles. Specifically, Fig. 4 reports the average accuracy over all video pairs with respect to the temporal segments. As can be seen, our technique outperforms all the other methods for all temporal phases with an average margin of $7-8\%$. Furthermore, the results illustrate once more that performing joint spatio-temporal alignemnt derives better results than applying the spatial and temporal alignment independently. Finally, we further evaluate the performance of the proposed method by applying different spatial alignment approaches (unsupervised, manual annotations, random initialisation), similar to MMI case. Due to limited space, this experiment is included in the supplementary material along with the CAD curves for each temporal phase separately as well as experiments in spatial alignment.

4. Conclusion

We proposed the first, to the best of our knowledge, spatio-temporal methodology for deformable face alignment. We proposed a novel component analysis for the task and we explored some of its theoretical properties, as well as its relationship with other component analysis (e.g., CCA). We showed that our methodology outperforms state-of-the-art temporal alignment methods that make use of manual image alignment. We also showed that it is advantageous to jointly solve the problems of spatial and temporal alignment than solving them independently.

5. Acknowledgements

The work of E. Antonakos was supported by EPSRC project EP/J017787/1 (4DFAB). The work of S. Zafeiriou was funded by the FiDiPro program of Tekes (project number: 1849/31/2015). The work of M. Pantic and L. Zafeiriou was partially supported by EPSRC project EP/N007743/1 (FACER2VM).

References

[1] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the ACM International Conference on Multimedia*, pages 679–682. ACM, 2014. 1

[2] E. Antonakos and S. Zafeiriou. Automatic construction of deformable models in-the-wild. In *CVPR*, pages 1813–1820, 2014. 1, 2

[3] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *IJCV*, 56(3):221–255, 2004. 1, 5

[4] D. P. Bertsekas. *Dynamic programming and optimal control*. Athena Scientific Belmont, MA, 1995. 3

[5] M. B. Blaschko, C. H. Lampert, and A. Gretton. Semi-supervised laplacian regularization of kernel canonical correlation analysis. In *Machine Learning and Knowledge Discovery in Databases*, pages 133–145. Springer, 2008. 5, 6

[6] V. N. Boddeti, T. Kanade, and B. V. Kumar. Correlation filters for object alignment. In *CVPR*, pages 2291–2298, 2013. 1

[7] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE T-PAMI*, 33(8):1548–1560, 2011. 5

[8] Y. Caspi and M. Irani. Spatio-temporal alignment of sequences. *IEEE T-PAMI*, 24(11):1409–1424, 2002. 1

[9] X. Cheng, S. Sridharan, J. Saragih, and S. Lucey. Rank minimization across appearance and shape for aam ensemble fitting. In *ICCV*, pages 577–584, 2013. 2

[10] F. De la Torre. A least-squares framework for component analysis. *IEEE T-PAMI*, 34(6):1041–1055, 2012. 3, 6

[11] H. Dibeklioglu, A. A. Salah, and T. Gevers. Uva-nemo smile database. http://www.uva-nemo.org/. 6

[12] F. Diego, J. Serrat, and A. M. Lopez. Joint spatio-temporal alignment of sequences. *IEEE T-MM*, 15(6):1377–1387, 2013. 1

[13] P. Ekman and W. V. Friesen. Facial action coding system. 1977. 3

[14] E. Hsu, K. Pulli, and J. Popović. Style translation for human motion. *ACM TOG*, 24(3):1082–1089, 2005. 1

[15] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez. View-independent action recognition from temporal self-similarities. *IEEE T-PAMI*, 33(1):172–185, 2011. 3

[16] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004. 1, 2, 6

[17] M. Nicolaou, V. Pavlovic, and M. Pantic. Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations. *IEEE T-PAMI*, 36(7):1299–1311, July 2014. 1, 2, 6, 8

[18] Y. Panagakis, M. Nicolaou, S. Zafeiriou, and M. Pantic. Robust correlated and individual component analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*. 4

[19] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic. Robust canonical time warping for the alignment of grossly corrupted sequences. In *CVPR*, pages 540–547, 2013. 1

[20] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *ICME*, pages 317–321, Amsterdam, The Netherlands, July 2005. 6

[21] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Raps: Robust and efficient automatic construction of person-specific deformable models. In *CVPR*, June 2014. 2

[22] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR'W*, Portland Oregon, USA, June 2013. 1

[23] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE T-PAMI*, 37(6):1113, 2015. 1

[24] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014. 1

[25] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *CVPR*, pages 1851–1858, 2014. 2

[26] M. F. Valstar and M. Pantic. Mmi facial expression database. http://www.mmifacedb.com/. 6

[27] L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002. 3

[28] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013. 1

[29] L. Zafeiriou, E. Antonakos, S. Zafeiriou, and M. Pantic. Joint unsupervised face alignment and behaviour analysis. In *ECCV*, pages 167–183. Springer, 2014. 2, 3, 4

[30] L. Zafeiriou, M. A. Nicolaou, S. Zafeiriou, S. Nikitidis, and M. Pantic. Learning slow features for behaviour analysis. In *ICCV*, 2013. 1, 2

[31] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24, 2015. 6

[32] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE T-PAMI*, 31(1):39–58, 2009. 8

[33] C. Zhao, W.-K. Cham, and X. Wang. Joint face alignment with a generic deformable face model. In *CVPR*, pages 561–568, 2011. 2

[34] F. Zhou and F. De la Torre. Canonical time warping for alignment of human behavior. In *NIPS*, pages 2286–2294, 2009. 1, 2, 3

[35] F. Zhou and F. De la Torre. Generalized time warping for multi-modal alignment of human motion. In *CVPR*, pages 1282–1289, 2012. 1

[36] F. Zhou, F. De la Torre, and J. K. Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE T-PAMI*, 35(3):582–596, 2013. 1

[37] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006. 4