Deep Canonical Time Warping for simultaneous alignment and representation learning of sequences

George Trigeorgis, Mihalis A. Nicolaou, *Member, IEEE,* Björn W. Schuller, *Senior member, IEEE*
Stefanos Zafeiriou, *Member, IEEE*

Abstract—Machine learning algorithms for the analysis of time-series often depend on the assumption that utilised data are temporally aligned. Any temporal discrepancies arising in the data is certain to lead to ill-generalisable models, which in turn fail to correctly capture properties of the task at hand. The temporal alignment of time-series is thus a crucial challenge manifesting in a multitude of applications. Nevertheless, the vast majority of algorithms oriented towards temporal alignment are either applied directly on the observation space or simply utilise linear projections - thus failing to capture complex, hierarchical non-linear representations that may prove beneficial, especially when dealing with multi-modal data (e.g., visual and acoustic information). To this end, we present Deep Canonical Time Warping (DCTW), a method that automatically learns non-linear representations of multiple time-series that are (i) maximally correlated in a shared subspace, and (ii) temporally aligned. Furthermore, we extend DCTW to a supervised setting, where during training, available labels can be utilised towards enhancing the alignment process. By means of experiments on four datasets, we show that the representations learnt significantly outperform state-of-the-art methods in temporal alignment, elegantly handling scenarios with heterogeneous feature sets, such as the temporal alignment of acoustic and visual information.

Index Terms—time warping, cca, lda, dcca, dda, deep learning, shared representations, dctw

1 Introduction

The alignment of multiple data sequences is a commonly arising problem, raised in multiple fields related to machine learning, such as signal, speech and audio analysis [33], computer vision [6], graphics [5] and bioinformatics [1]. Example applications range from the temporal alignment of facial expressions and motion capture data [43], [44], to the alignment for human action recognition [40], and speech [22].

The most prominent temporal alignment method is Dynamic Time Warping (DTW) [33], which identifies the optimal warping path that minimises the Euclidean distance between two time-series. While DTW has found wide application over the past decades, the application is limited mainly due to the inherent inability of DTW to handle observations of different or high dimensionality since it directly operates on the observation space. Motivated by this limitation while recognising that this scenario is commonly encountered in real-world applications (*e.g.*, capturing data from multiple sensors), in [43] an extension to DTW is proposed. Coined Canonical Time Warping (CTW), the method combines Canonical Correlation Analysis (CCA) and DTW by aligning the two sequences in a common, latent subspace of reduced dimensionality whereon the two sequences are maximally correlated. Other extensions of DTW include the integration of manifold learning, thus facilitating the alignment of sequences lying on different manifolds [15], [40] while in [36], [44] constraints are introduced in order to guarantee monotonicity and adaptively constrain the temporal warping. It should be noted that in [44], a multi-set variant of CCA is utilised [18] thus enabling the temporal alignment of multiple sequences, while a Gauss-Newton temporal warping method is proposed.

While methods aimed at solving the problem of temporal alignment have been successful in a wide spectrum of applications, most of the aforementioned techniques find a single *linear* projection for each sequence. While this may suffice for certain problem classes, in many real world applications the data are likely to be embedded with more complex, possibly hierarchical and non-linear structures. A prominent example lies in the alignment of non-linear acoustic features with raw pixels extracted from a video stream (for instance, in the audiovisual analysis of speech, where the temporal misalignment is a common problem). The mapping between these modalities is deemed highly nonlinear, and in order to appropriately align them in time this needs to be taken into account. An approach towards extracting such complex non-linear transformations is via adopting the principles associated with the recent revival of deep neural network architectural models. Such architectures have been successfully applied in a multitude of problems, including feature extraction and dimensionality reduction [20], feature extraction for object recognition and detection [14], [25], feature extraction for face recognition [37], acoustic modelling in speech recognition [19], as well

- *G. Trigeorgis, S. Zafeiriou, and B. W. Schuller are with the Department of Computing, Imperial College London, SW7 2RH, London, UK*
  E-mail: g.trigeorgis@imperial.ac.uk
- *Mihalis A. Nicolaou is with the Department of Computing at Goldsmiths, University of London*
  E-mail: m.nicolaou@gold.ac.uk
- *Stefanos Zafeiriou is also with Center for Machine Vision and Signal Analysis, University of Oulu, Finland.*

as for extracting non-linear correlated features [2].

Of interest to us is also work that has evolved around multimodal learning. Specifically, deep architectures deemed very promising in several areas, often overcoming by a large margin traditionally used methods in various emotion and speech recognition tasks [24], [29], and on robotics applications with visual and depth data [41].

In this light, we propose Deep Canonical Time Warping (DCTW), a novel method aimed towards the alignment of multiple sequences that discovers complex, hierarchical representations which are both maximally correlated *and* temporally aligned. To the best of our knowledge, this work presents the *first* deep approach towards solving the problem of temporal alignment[1], which in addition offers very good scaling when dealing with large amounts of data. In more detail, this paper carries the following contributions: (i) we extend DTW-based temporal alignment methods to handle heterogeneous collections of features that may be connected via non-linear hierarchical mappings, (ii) in the process, we extend DCCA to (a) handle arbitrary temporal discrepancies in the observations and (b) cope with multiple (more than two) sequences, while (iii) we extend DCCA and DCTW in order to extract hierarchical, non-linear features in the presence of labelled data, thus enriched with *discriminative* properties. In order to do so, we exploit the optimisation problem of DCCA in order to provide a deep counterpart of Linear Discriminant Analysis (LDA), that is subsequently extend with time-warpings. We evaluate the proposed methods on a multitude of real data sets, where the performance gain in contrast to other state-of-the-art methods becomes clear.

The remainder of this paper is organised as follows. We firstly introduce related work in Section 2, while the proposed Deep Canonical Time Warping (DCTW) is presented in Section 3. In Section 4, we introduce supervision by presenting the Deep Discriminant Analysis (DDA) variant, along with the extension an extension that incorporates time warpings (DDATW). Finally, experimental results on several real datasets are presented in Section 5.

2 Related Work

2.1 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a shared-space component analysis method, that given two data matrices $\mathbf{X}_1, \mathbf{X}_2$ where $\mathbf{X}_i \in \mathbb{R}^{d_i \times T}$ recovers the loadings $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times d}$, $\mathbf{W}_2 \in \mathbb{R}^{d_2 \times d}$ that linearly project the data on a subspace where the linear correlation is maximised. This can be interpreted as discovering the shared information conveyed by all the datasets (or views). The correlation $\rho = \text{corr}(\mathbf{Y}_1, \mathbf{Y}_2)$ in the projected space $\mathbf{Y}_i = \mathbf{W}_i^\top \mathbf{X}_i$ can be written as

$$\boldsymbol{\rho} = \frac{\mathbb{E}[\mathbf{Y}_1 \mathbf{Y}_2^\top]}{\sqrt{\mathbb{E}[\mathbf{Y}_1 \mathbf{Y}_1^\top \mathbf{Y}_2 \mathbf{Y}_2^\top]}} \tag{1}$$

$$= \frac{\mathbf{W}_1^\top \mathbb{E}[\mathbf{X}_1 \mathbf{X}_2^\top] \mathbf{W}_2}{\sqrt{\mathbf{W}_1^\top \mathbb{E}[\mathbf{X}_1 \mathbf{X}_1^\top] \mathbf{W}_1 \mathbf{W}_2^\top \mathbb{E}[\mathbf{X}_2 \mathbf{X}_2^\top] \mathbf{W}_2}} \tag{2}$$

$$= \frac{\mathbf{W}_1^\top \boldsymbol{\Sigma}_{12} \mathbf{W}_2}{\sqrt{\mathbf{W}_1^\top \boldsymbol{\Sigma}_{11} \mathbf{W}_1 \mathbf{W}_2^\top \boldsymbol{\Sigma}_{22} \mathbf{W}_2}} \tag{3}$$

where $\boldsymbol{\Sigma}_{ij}$ denotes the empirical covariance between data matrices $\mathbf{X}_i$ and $\mathbf{X}_j$[2]. There are multiple equivalent optimisation problems for discovering the optimal loadings $\mathbf{W}_i$ which maximise Equation 3 [9]. For instance, CCA can be formulated as a least-squares problem,

$$\begin{aligned} &\underset{\mathbf{W}_1, \mathbf{W}_2}{\arg\min} \ \|\mathbf{W}_1^\top \mathbf{X}_1 - \mathbf{W}_2^\top \mathbf{X}_2\|_F^2 \\ &\text{subject to: } \mathbf{W}_1^\top \mathbf{X}_1 \mathbf{X}_1^\top \mathbf{W}_1 = \mathbf{I}, \\ &\qquad\qquad \mathbf{W}_2^\top \mathbf{X}_2 \mathbf{X}_2^\top \mathbf{W}_2 = \mathbf{I}, \end{aligned} \tag{4}$$

and as

$$\begin{aligned} &\underset{\mathbf{W}_1, \mathbf{W}_2}{\arg\min} \ \|\mathbf{W}_1^\top \mathbf{X}_1 - \mathbf{W}_2^\top \mathbf{X}_2\|_F^2 \\ &= \underset{\mathbf{W}_1, \mathbf{W}_2}{\arg\min} \ \text{tr}\left(\mathbf{W}_1^\top \boldsymbol{\Sigma}_1 \mathbf{W}_1 - 2\mathbf{W}_1^\top \boldsymbol{\Sigma}_{12} \mathbf{W}_2 + \mathbf{W}_2^\top \boldsymbol{\Sigma}_2 \mathbf{W}_2\right) \end{aligned}$$

we can reformulate this as a trace optimisation problem as the projected covariance terms $\mathbf{W}_1^\top \boldsymbol{\Sigma}_1 \mathbf{W}_1$ and $\mathbf{W}_2^\top \boldsymbol{\Sigma}_2 \mathbf{W}_2$ are substituted due to the orthogonality constraints with an identity matrix.

$$\begin{aligned} &\underset{\mathbf{W}_1, \mathbf{W}_2}{\arg\max} \ \text{tr}\left(\mathbf{W}_1^\top \mathbf{X}_1 \mathbf{X}_2^\top \mathbf{W}_2\right) \\ &\text{subject to } \mathbf{W}_1^\top \mathbf{X}_1 \mathbf{X}_1^\top \mathbf{W}_1 = \mathbf{I}, \\ &\qquad\qquad \mathbf{W}_2^\top \mathbf{X}_2 \mathbf{X}_2^\top \mathbf{W}_2 = \mathbf{I}, \end{aligned} \tag{5}$$

where in both cases we exploit the scale invariance of the correlation coefficient with respect to the loadings in the constraints. The solution in both cases is given by the eigenvectors corresponding to the d largest eigenvalues of the generalised eigenvalue problem

$$\begin{pmatrix} \mathbf{0} & \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix} \boldsymbol{\Lambda}. \tag{6}$$

The eigenvalue problem can be also made symmetric by introducing $\mathbf{W}_1 = \boldsymbol{\Sigma}_{11}^{-\frac{1}{2}} \mathbf{V}_1$ and $\mathbf{W}_2 = \boldsymbol{\Sigma}_{22}^{-\frac{1}{2}} \mathbf{V}_2$.

$$\begin{pmatrix} \mathbf{0} & \boldsymbol{\Sigma}_{11}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-\frac{1}{2}} \\ \boldsymbol{\Sigma}_{22}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-\frac{1}{2}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} \boldsymbol{\Lambda}. \tag{7}$$

Note that an equivalent solution is obtained by resorting to Singular Value Decomposition (SVD) on the matrix $\mathbf{K} = \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1/2}$ [4], [27]. The optimal objective value of Equation 5 is then the sum of the largest d singular values of $\mathbf{K}$, while the optimal loadings are found by setting $\mathbf{W}_1 = \boldsymbol{\Sigma}_{11}^{-1/2} \mathbf{U}_d$ and $\mathbf{W}_2 = \boldsymbol{\Sigma}_{22}^{-1/2} \mathbf{V}_d$, with $\mathbf{U}_d$ and $\mathbf{V}_d$ being the left and right singular vectors of $\mathbf{K}$. Note that this interpretation is completely analogous to solving the corresponding generalised eigenvalue problem arising in Equation 7 and keeping the top d eigenvectors corresponding to the largest eigenvalues.

In the case of multiple sets of datasets, Multi-set CCA (MCCA) has been proposed [18], [30]. As expected the

1. A preliminary version of our work has appeared in [38].

2. Note that we assume zero-mean data to avoid cluttering the notation.

optimisation goal in this case then becomes to maximise the pairwise correlation scores of the m different data sets, subject to the orthogonality constraints.

$$\begin{aligned} \underset{\mathbf{W}_1,\ldots,\mathbf{W}_m}{\arg\min} & \sum_{i,j=1}^{m} \|\mathbf{W}_i^\top \mathbf{X}_i - \mathbf{W}_j^\top \mathbf{X}_j\|_F^2 \\ \text{subject to: } & \mathbf{W}_1^\top \mathbf{X}_1 \mathbf{X}_1^\top \mathbf{W}_1 = \mathbf{I}, \\ & \mathbf{W}_2^\top \mathbf{X}_2 \mathbf{X}_2^\top \mathbf{W}_2 = \mathbf{I}, \\ & \vdots \\ & \mathbf{W}_m^\top \mathbf{X}_m \mathbf{X}_m^\top \mathbf{W}_m = \mathbf{I}. \end{aligned} \tag{8}$$

Recently, in order to facilitate the extraction of non-linear correlated transformations, a methodology inspired by CCA called Deep CCA (DCCA) [2] was proposed. In more detail, motivated by the recent success of deep architectures, DCCA assumes a network of multiple stacked layers consisting of nonlinear transformations for each data set i, with parameters $\theta_i = \{\theta_i^1, ..., \theta_i^l\}$, where l is the number of layers. Assuming the transformation applied by the network corresponding to data set i is represented as $f_i(\mathbf{X}_i; \theta_i)$, the optimal parameters are found by solving

$$\underset{\theta_1,\theta_2}{\arg\max}\ \text{corr}(f_1(\mathbf{X}_1;\theta_1), f_2(\mathbf{X}_2;\theta_2)). \tag{9}$$

Let us assume that in each of the networks, the final layer has d maximally correlated units in an analogous fashion to the classical CCA Equation 3. In particular, we consider that $\tilde{\mathbf{X}}_i$ denotes the transformed input data sets, $\tilde{\mathbf{X}}_i = f_i(\mathbf{X}_i; \theta_i)$ and that the covariances $\tilde{\mathbf{\Sigma}}_{ij}$ are now estimated on $\tilde{\mathbf{X}}$, *i.e.*, $\tilde{\mathbf{\Sigma}}_{ii} = \frac{1}{T-1}\tilde{\mathbf{X}}_i(\mathbf{I} - \frac{1}{T}\mathbf{1}\mathbf{1}^\top)\tilde{\mathbf{X}}_i^\top$, where T is the length of the sequence $\mathbf{X}_i$. As described above for classical CCA (Equation 5), the optimal objective value is the sum of the k largest singular values of $\mathbf{K} = \tilde{\mathbf{\Sigma}}_{11}^{-1/2}\tilde{\mathbf{\Sigma}}_{12}\tilde{\mathbf{\Sigma}}_{22}^{-1/2}$, which is exactly the nuclear norm of $\mathbf{K}$, $\|\mathbf{K}\|_* = \text{trace}(\sqrt{\mathbf{K}\mathbf{K}^\top})$. Problem 9 now becomes

$$\underset{\theta_1,\theta_2}{\arg\max}\ \|\mathbf{K}\|_* . \tag{10}$$

and this is precisely the loss function that is backpropagated through the network[3] [2]. Put simply, the networks are optimised towards producing features which exhibit high canonical correlation coefficients.

2.2 Time Warping

Given two data matrices $\mathbf{X}_1 \in \mathbb{R}^{d\times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{d\times T_2}$ Dynamic Time Warping (DTW) aims to eliminate temporal discrepancies arising in the data by optimising Equation 11,

$$\begin{aligned} \underset{\mathbf{\Delta}_1,\mathbf{\Delta}_2}{\arg\min}\ & \|\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{X}_2\mathbf{\Delta}_2\|_F^2 \\ \text{subject to: } & \mathbf{\Delta}_1 \in \{0,1\}^{T_1\times T}, \\ & \mathbf{\Delta}_2 \in \{0,1\}^{T_2\times T}, \end{aligned} \tag{11}$$

where $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$ are binary selection matrices [43] that encode the alignment path, effectively remapping the the samples of each sequence to a common temporal scale. Although the number of plausible alignment paths is exponential with respect to T_1T_2, by employing dynamic programming, DTW infers the optimal alignment path (in terms of Equation 11) in $\mathcal{O}(T_1T_2)$. Finally, the DTW solution satisfies the boundary, continuity, and monotonicity constraints [33].

The main limitation of DTW lies in the inherent inability to handle sequences of varying feature dimensionality, which is commonly the case when examining data acquired from multiple sensors. Furthermore, DTW is prone to failure when one or more sequences are perturbed by arbitrary affine transformations. To this end, the Canonical Time Warping (CTW) [43] elegantly combines the least-squares formulations of DTW (Equation 11) and CCA (Equation 4), thus facilitating the utilisation of sequences with varying dimensionalities, while simultaneously performing feature selection and temporal alignment. In more detail, given $\mathbf{X}_1 \in \mathbb{R}^{d_1\times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{d_2\times T_2}$, the CTW problem is posed as

$$\begin{aligned} \underset{\mathbf{W}_1,\mathbf{W}_2,\mathbf{\Delta}_1,\mathbf{\Delta}_2}{\arg\min}\ & \|\mathbf{W}_1^\top\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{W}_2^\top\mathbf{X}_2\mathbf{\Delta}_2\|_F^2 \\ \text{subject to: } & \mathbf{W}_1^\top\mathbf{X}_1\mathbf{\Delta}_1\mathbf{\Delta}_1^\top\mathbf{X}_1^\top\mathbf{W}_1 = \mathbf{I}, \\ & \mathbf{W}_2^\top\mathbf{X}_2\mathbf{\Delta}_2\mathbf{\Delta}_2^\top\mathbf{X}_2^\top\mathbf{W}_2 = \mathbf{I}, \\ & \mathbf{W}_1^\top\mathbf{X}_1\mathbf{\Delta}_1\mathbf{\Delta}_2^\top\mathbf{X}_2^\top\mathbf{W}_2 = \mathbf{D}, \\ & \mathbf{X}_1\mathbf{\Delta}_1\mathbf{1} = \mathbf{X}_2\mathbf{\Delta}_2\mathbf{1} = \mathbf{0} \\ & \mathbf{\Delta}_1 \in \{0,1\}^{T_1\times T}, \mathbf{\Delta}_2 \in \{0,1\}^{T_2\times T}, \end{aligned} \tag{12}$$

where the loadings $\mathbf{W}_1 \in \mathbb{R}^{d\times T_1}$ and $\mathbf{W}_2 \in \mathbb{R}^{d\times T_2}$ project the observations onto a reduced dimensionality subspace where they are maximally linearly correlated, $\mathbf{D}$ is a diagonal matrix and $\mathbf{1}$ is a vector of all 1's of appropriate dimensions. The constraints in Equation 12, mostly inherited by CCA, deem the CTW solution translation, rotation, and scaling invariant. We note that the final solution is obtained by alternating between solving CCA (by fixing $\mathbf{X}_i\mathbf{\Delta}_i$) and DTW (by fixing $\mathbf{W}_i^\top\mathbf{X}_i$).

3 Deep Canonical Time Warping (DCTW)

The goal of Deep Canonical Time Warping (DCTW) is to discover a hierarchical non-linear representation of the data sets $\mathbf{X}_i, i = \{1, 2\}$ where the transformed features are (i) temporally aligned with each other, and (ii) maximally correlated. To this end, let us consider that $f_i(\mathbf{X}_i; \theta_i)$ represents the final layer activations of the corresponding network for dataset $\mathbf{X}_i$[4]. We propose to optimise the following objective,

$$\begin{aligned} \underset{\theta_1,\theta_2,\mathbf{\Delta}_1,\mathbf{\Delta}_2}{\arg\min}\ & \|f_1(\mathbf{X}_1;\theta_1)\mathbf{\Delta}_1 - f_2(\mathbf{X}_2;\theta_2)\mathbf{\Delta}_2\|_F^2 \\ \text{subject to: } & f_1(\mathbf{X}_1;\theta_1)\mathbf{\Delta}_1\mathbf{\Delta}_1^\top f_1(\mathbf{X}_1;\theta_1)^\top = \mathbf{I}, \\ & f_2(\mathbf{X}_2;\theta_2)\mathbf{\Delta}_2\mathbf{\Delta}_2^\top f_2(\mathbf{X}_2;\theta_2)^\top = \mathbf{I}, \\ & f_1(\mathbf{X}_1;\theta_1)\mathbf{\Delta}_1\mathbf{\Delta}_2^\top f_2(\mathbf{X}_2;\theta_2) = \mathbf{D}, \\ & f_1^p(\mathbf{X}_1;\theta_1)\mathbf{\Delta}_1\mathbf{1} = f_2^p(\mathbf{X}_2;\theta_2)\mathbf{\Delta}_2\mathbf{1} = \mathbf{0}, \\ & \mathbf{\Delta}_1 \in \{0,1\}^{T_1\times T}, \mathbf{\Delta}_2 \in \{0,1\}^{T_2\times T} \end{aligned} \tag{13}$$

where as defined for Equation 12, $\mathbf{D}$ is a diagonal matrix and $\mathbf{1}$ is an appropriate dimensionality vector of all 1's.

3. Since the nuclear norm is non-differentiable RWand motivated by [3], in [2] the subgradient of the nuclear norm is utilised in gradient descent.

4. We denote the penultimate layer of the network as $f_i^p(\mathbf{X}_i; \theta_i)$ which is then followed by a linear layer.

Clearly, the objective can be solved via alternating optimisation. Given the activation of the output nodes of each network i, DTW recovers the optimal warping matrices $\mathbf{\Delta}_i$ which temporally align them. Nevertheless, the inverse is not so straight-forward, since we have no closed form solution for finding the optimal non-linear stacked transformation applied by the network. We therefore resort to finding the optimal parameters of each network by utilising backpropagation. Having discovered the warping matrices $\mathbf{\Delta}_i$, the problem becomes equivalent to applying a variant of DCCA in order to infer the maximally correlated non-linear transformation on the temporally aligned input features. This requires that the covariances are reformulated as $\hat{\mathbf{\Sigma}}_{ij} = \frac{1}{T-1} f_i(\mathbf{X}_i;\theta_i)\mathbf{\Delta}_i \mathbf{C}_T \mathbf{\Delta}_j^\top f_j(\mathbf{X}_j;\theta_j)^\top$, where $\mathbf{C}_T$ is the centering matrix, $\mathbf{C}_T = \mathbf{I} - \frac{1}{T}\mathbf{1}\mathbf{1}^\top$. By defining $\mathbf{K}_{\mathcal{DCTW}} = \hat{\mathbf{\Sigma}}_{11}^{-1/2}\hat{\mathbf{\Sigma}}_{12}\hat{\mathbf{\Sigma}}_{22}^{-1/2}$, we now have that

$$\text{corr}(f_1(\mathbf{X}_1;\theta_1)\mathbf{\Delta}_1, f_2(\mathbf{X}_2;\theta_2)\mathbf{\Delta}_2) = \|\mathbf{K}_{\mathcal{DCTW}}\|_*. \quad (14)$$

We optimise this quantity in a gradient-ascent fashion by utilising the subgradient of Equation 14 [3], since the gradient can not be computed analytically. By assuming that $\mathbf{Y}_i = f_i(\mathbf{X}_i;\theta_i)$ for each of network i and $\mathbf{U}\mathbf{S}\mathbf{V}^\top = \mathbf{K}_{\mathcal{DCTW}}$ is the singular value decomposition of $\mathbf{K}_{\mathcal{DCTW}}$, then the subgradient for the last layer is defined as

$$\begin{aligned}
\mathbf{F}^{(\text{pos})} &= \hat{\mathbf{\Sigma}}_{11}^{-1/2}\mathbf{U}\mathbf{V}^\top\hat{\mathbf{\Sigma}}_{22}^{-1/2}\mathbf{Y}_2\mathbf{\Delta}_2\mathbf{C}_T \\
\mathbf{F}^{(\text{neg})} &= \hat{\mathbf{\Sigma}}_{11}^{-1/2}\mathbf{U}\mathbf{S}\mathbf{U}^\top\hat{\mathbf{\Sigma}}_{11}^{-1/2}\mathbf{Y}_1\mathbf{\Delta}_1\mathbf{C}_T \\
\frac{\partial \|\mathbf{K}_{\mathcal{DCTW}}\|_*}{\partial \mathbf{Y}_1} &= \frac{1}{T-1}\left(\mathbf{F}^{(\text{pos})} - \mathbf{F}^{(\text{neg})}\right). \quad (15)
\end{aligned}$$

At this point, it is clear that CTW is a special case of DCTW. In fact, we arrive at CTW (subsection 2.2) by simply considering a network with one layer. In this case, by setting $f_i(\mathbf{X}_i;\theta_i) = \mathbf{W}_i^\top\mathbf{X}_i$, Equation 17 becomes equivalent to Equation 12, while solving Equation 14 by means of Singular Value Decomposition (SVD) on $\mathbf{K}_{\mathcal{DCTW}}$ provides equivalent loadings to the ones obtained by CTW via eigenanalysis.

Finally, we note that we can easily extend DCTW to handle multiple (more than 2) data sets, by incorporating a similar objective to the Multi-set Canonical Correlation Analysis (MCCA) [18], [30]. In more detail, instead of Equation 14 we now optimise

$$\begin{aligned}
&\sum_{i,j=1}^{m} \text{corr}(f_i(\mathbf{X}_i;\theta_i)\mathbf{\Delta}_i, f_j(\mathbf{X}_j;\theta_j)\mathbf{\Delta}_j) \\
&= \sum_{i,j}^{m} \left\|\mathbf{K}_{\mathcal{DCTW}}^{ij}\right\|_* \quad (16)
\end{aligned}$$

where m is the number of sequences and $\mathbf{K}_{\mathcal{DCTW}}^{ij} = \hat{\mathbf{\Sigma}}_{ii}^{-1/2}\hat{\mathbf{\Sigma}}_{ij}\hat{\mathbf{\Sigma}}_{jj}^{-1/2}$. This leads to the following optimisation problem,

$$\begin{aligned}
&\underset{\forall k.\theta_k, \mathbf{\Delta}_k}{\arg\min} \sum_{i,j=1}^{m} \|f_i(\mathbf{X}_i;\theta_i)\mathbf{\Delta}_i - f_j(\mathbf{X}_j;\theta_j)\mathbf{\Delta}_j\|_F^2 \\
&\text{subject to: } \forall k. f_k(\mathbf{X}_k;\theta_k)\mathbf{\Delta}_k\mathbf{\Delta}_k^\top f_k(\mathbf{X}_k;\theta_k)^\top = \mathbf{I}, \\
&\qquad \forall i,j. f_i(\mathbf{X}_i;\theta_i)\mathbf{\Delta}_i\mathbf{\Delta}_j^\top f_j(\mathbf{X}_j;\theta_j) = \mathbf{D}, \\
&\qquad \forall k. f_k^p(\mathbf{X}_k;\theta_k)\mathbf{\Delta}_k\mathbf{1} = \mathbf{0}, \\
&\qquad \forall k. \mathbf{\Delta}_k \in \{0,1\}^{T_k \times T} \quad (17)
\end{aligned}$$

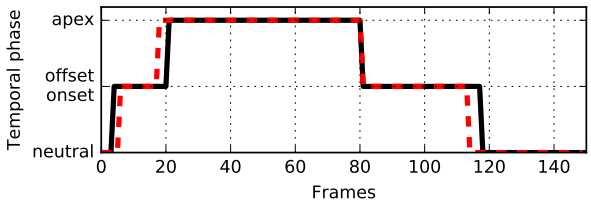

Fig. 2. The ground-truth temporal segments (- - -) and the corresponding predicted temporal phases (——) for each of the frames of a video displaying AU12 using DDATW.

The subgradient of Equation 16 then becomes

$$\begin{aligned}
&\frac{\partial \sum_{i,j}^{m} \left\|\mathbf{K}_{\mathcal{DCTW}}^{ij}\right\|_*}{\partial \mathbf{Y}_i} \\
&= \sum_{j}^{m} \frac{\partial \left\|\mathbf{K}_{\mathcal{DCTW}}^{ij}\right\|_*}{\partial \mathbf{Y}_i} + \sum_{j}^{m} \frac{\partial \left\|\mathbf{K}_{\mathcal{DCTW}}^{ji}\right\|_*}{\partial \mathbf{Y}_i} \\
&= 2\sum_{j}^{m} \frac{\partial \left\|\mathbf{K}_{\mathcal{DCTW}}^{ij}\right\|_*}{\partial \mathbf{Y}_i}. \quad (18)
\end{aligned}$$

Note that by setting $\mathbf{\Delta}_i = \mathbf{I}$, Equation 16 becomes an objective for learning transformations for multiple sequences via DCCA [2]. Finally, we note that any warping method can be used in place of DTW for inferring the warping matrices $\mathbf{\Delta}_i$ (e.g., [44]), while DCTW is further illustrated in Figure 1.

3.1 Topology

At this point we should clarify that our model is topology-agnostic; our cost-function is optimised regardless of the number of layers or neuron type. Although we experimentally show later on that a 3-layer network can be sufficient, more elaborated topologies can be used that better suit the task-at-hand. An obvious example for this would be the problem of learning the optimal alignment and time-invariant representations of visual modalities such as videos. In this case, to reduce the free parameters of the model, convolutional neurons can be employed, and moreover the parameters for each network f_i for $0 < i < m$ can be tied (see Siamese networks [7]).

4 Supervised Deep Time Warping

The deep time-warping approach described in Section 3 recovers the appropriate non-linear transformations for temporally aligning a set of arbitrary sequences (*e.g.*, temporally aligning videos of subjects performing the same, or similar, facial expression). This is done by optimizing an appropriate loss function (Equation 17). Nevertheless, in many similar problem settings, a set of labels characterising the temporal information contained in the sequences is readily available (*e.g.*, labels containing the temporal phase of facial Action Units activated in the video). Although such labels can be readily utilised in order to evaluate the resulting alignment, this information remains unexploited in DCTW, as well as in other state-of-the-art time-warping methods such as [40], [43], [45].

In this section, we exploit the flexibility of the optimisation problem proposed for DCTW in order to exploit

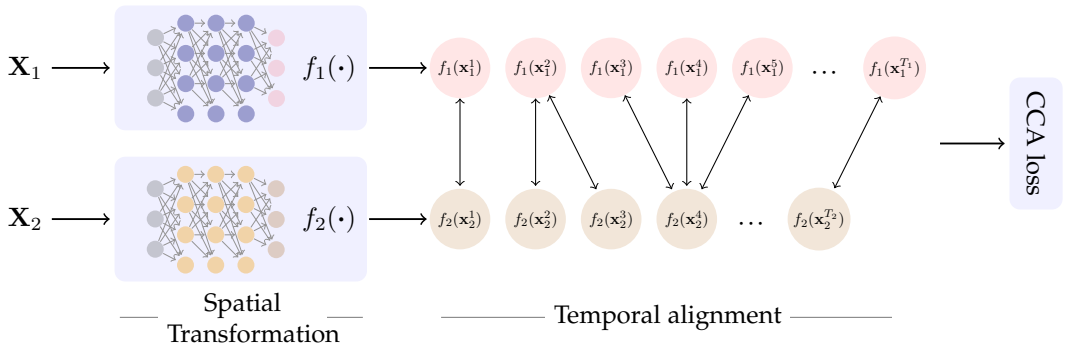

Fig. 1. Illustration of the DCTW architecture with two networks, one for each temporal sequence. The model is trained end-to-end, first performing a spatial transformation of the data samples and then a temporal transformation such as the temporal sequences are maximally correlated.

labelled information with the goal of enhancing performance on unseen, unlabelled data. By considering the setting where the sequences at-hand are annotated with discrete labels corresponding to particular temporal events, we firstly show that by appropriately modifying the objective for DCCA, we arrive at a numerically stable, non-linear variant of the traditionally used Linear Discriminant Analysis (LDA), which we call Deep Discriminant Analysis (DDA). While DDA can be straightforwardly applied within a general supervised learning context in order to learn non-linear discriminative transformations, we subsequently extend the proposed optimisation problem for DCTW (Equation 17) by incorporating time-warping in the objective function. This leads to the Deep Discriminant Analysis with Time Warping (DDATW) method, that can be utilised towards temporally aligning multiple sequences while exploiting label information.

4.1 Deep Discriminant Analysis

Let us assume a set of T samples $\mathbf{x}_i$ is given, with a label $y_i \in \{1, \dots, C\}$ corresponding to each sample. The classical Linear Discriminant Analysis (LDA) [13] computes a linear transformation of $\mathbf{W}$ that maximises the dispersion of class means while minimising the within-class variance. A standard formulation of LDA is given by the following a trace optimisation problem,

$$\begin{aligned} \arg\max_{\mathbf{W}} \quad & \text{tr}(\mathbf{W}^\top \mathbf{S_b} \mathbf{W}) \\ & \text{s.t. } \mathbf{W}^\top \mathbf{S_t} \mathbf{W} = \mathbf{I} \end{aligned} \tag{19}$$

where $\mathbf{S_b} = \sum_{i=C} n_i \mathbf{m}_i \mathbf{m}_i^\top$, n_i are the number of samples in i-th class and $\mathbf{m}_i = \frac{1}{n_i} \sum_{y_k=i} \mathbf{x}_k$ the corresponding mean. Furthermore, $\mathbf{S_t} = \mathbf{X}\mathbf{X}^\top$ is the total scatter matrix.

In matrix notation the between class scatter matrix can be constructed as follows,

$$\mathbf{S_b} = \mathbf{X}\mathbf{G}(\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top \mathbf{X}^\top$$

where $\mathbf{G} \in \mathbb{R}^{n \times c}$ is an indicator matrix in which $\sum_j g_{ij} = 1,\ g_{ij} \in \{0, 1\}$, and g_{ij} is 1 iff data sample i belongs to class j, and 0 otherwise. Thus $\mathbf{XG}$ is a matrix of the group sums, and $\mathbf{XG}(\mathbf{G}^\top \mathbf{G})^{-1}$ is a matrix which weights the sums with the respective number of data samples of each class.

The theory developed in [10] showed that there is an equivalence between least-square and trace optimisation problems. In particular, the problem of finding the optimal $\mathbf{W}$ that maps the data to labels can be written as

$$\arg\min_{\mathbf{W}} \|(\mathbf{G}^\top \mathbf{G})^{-\frac{1}{2}} (\mathbf{G}^\top - \mathbf{W}^\top \mathbf{X})\|_F^2$$

The above is equivalent to finding the optimal $\mathbf{W}$ from the following trace optimisation problem

$$\begin{aligned} \arg\max_{\mathbf{W}} \ & \text{tr}[\mathbf{W}^\top \mathbf{X}\mathbf{G}(\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top \mathbf{X}^\top \mathbf{W}] \\ & \text{s.t. } \mathbf{W}^\top \mathbf{X}\mathbf{X}^\top \mathbf{W} = \mathbf{I}, \end{aligned} \tag{20}$$

which is precisely the problem formulation for LDA (Equation 19).

As the connection between CCA (Equation 5) and LDA (Equation 20) is now established, we can easily extend LDA to a non-linear, hierarchical discriminant counterpart by taking advantage of the DCCA problem formulation in Equation 10. In more detail, the optimisation problem for Deep Discriminant Analysis (DDA) can be formulated as

$$\begin{aligned} \arg\min_{\theta} \ & \|f(\mathbf{X}; \theta) - (\mathbf{G}^\top \mathbf{G})^{-\frac{1}{2}} \mathbf{G}^\top\|_F^2 \\ \text{subject to: } & f(\mathbf{X}; \theta) f(\mathbf{X}; \theta)^\top = \mathbf{I}, \\ & f^p(\mathbf{X}; \theta) \mathbf{1} = \mathbf{0}, \end{aligned}$$

or equivalently by using the trace norm formulation as

$$\arg\max_{\theta} \|\mathbf{K}_{\text{LDA}}\|_* , \tag{21}$$

where $\mathbf{K}_{\text{LDA}} = \tilde{\mathbf{\Sigma}}_{11}^{\text{LDA}-1/2} \tilde{\mathbf{\Sigma}}_{12}^{\text{LDA}} \tilde{\mathbf{\Sigma}}_{22}^{\text{LDA}-1/2}$, $\tilde{\mathbf{\Sigma}}_{12}^{\text{LDA}} = \frac{1}{T-1} \tilde{\mathbf{X}}_i \mathbf{C}_T \mathbf{G} (\mathbf{G}^\top \mathbf{G})^{-\frac{1}{2}}$, $\tilde{\mathbf{\Sigma}}_{22}^{\text{LDA}} = \mathbf{G}(\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top$, $\tilde{\mathbf{\Sigma}}_{11}^{\text{LDA}} = \frac{1}{T-1} \tilde{\mathbf{X}}_i \mathbf{C}_T \tilde{\mathbf{X}}_i^\top$ with $\tilde{\mathbf{X}}_i = f_i(\mathbf{X}_i; \theta_i)$, while $\mathbf{C}_T = \mathbf{I} - \frac{1}{T} \mathbf{1}\mathbf{1}^\top$ denotes the centring matrix. We note that a Deep Linear Discriminant Analysis method has been recently proposed in [11], using a direct application of the LDA optimisation problem based on covariance diagonalisation, an approach that the authors found to be quite numerically unstable. On the contrary, the proposed DDA transformations based on (Equation 21) are found in a similar manner as DCCA, that is by using the sub-gradients of the nuclear norm, a process that involves computing the SVD. This approach can be more stable since the SVD decomposition exists for any matrix, not just for matrices that can be diagonalised.

4.2 Deep Discriminant Analysis with Time Warpings

The Deep Discriminant Analysis (DDA) method proposed in the previous section involves the optimisation of a trace norm in a similar manner to DCCA. DDA can thus be extended to incorporate time warpings, resulting in the proposed Deep Discriminant Analysis with Time Warpings (DDATW). That is, we can incorporate warpings by simply replacing $\tilde{\mathbf{X}}_i$ with $\tilde{\mathbf{X}}_i\boldsymbol{\Delta}_i$ and $(\mathbf{G}^\top\mathbf{G})^{-\frac{1}{2}}\mathbf{G}^\top$ with $(\mathbf{G}^\top\mathbf{G})^{-\frac{1}{2}}\mathbf{G}^\top\boldsymbol{\Delta}_g$ in Equation 21. In essence, we are solving an equivalent problem to the one described in Equation 14, namely

$$\begin{aligned}\underset{\forall k.\theta_k,\boldsymbol{\Delta}_k}{\arg\min} & \sum_{i,j=1}^{m} \|f_i(\mathbf{X}_i;\theta_i)\boldsymbol{\Delta}_i - (\mathbf{G}^\top\mathbf{G})^{-\frac{1}{2}}\mathbf{G}^\top\boldsymbol{\Delta}_j\|_F^2 \\ \text{subject to: } & \forall k. f_k(\mathbf{X}_k;\theta_k)\boldsymbol{\Delta}_k\boldsymbol{\Delta}_k^\top f_k(\mathbf{X}_k;\theta_k)^\top = \mathbf{I}, \\ & \forall i,j. f_i(\mathbf{X}_i;\theta_i)\boldsymbol{\Delta}_i\boldsymbol{\Delta}_j^\top f_j(\mathbf{X}_j;\theta_j) = \mathbf{D}, \\ & \forall k. f_k^p(\mathbf{X}_k;\theta_k)\boldsymbol{\Delta}_k\mathbf{1} = \mathbf{0}, \\ & \forall k. \boldsymbol{\Delta}_k \in \{0,1\}^{T_k\times T}. \end{aligned} \quad (22)$$

This formulation becomes particularly useful in cases when tackling tasks where discrete, temporal labels are available, for example, in case of annotating the temporal segments of the activations of facial Action Units (AUs). In particular, since in the vast majority of cases such labels are obtained by manually annotating the videos at hand, it is likely that artifacts such as lags and misalignments between labels and features may arise (e.g., an annotation that indicates that a particular AU has reached the apex phase after the actual phase has been actually reached in the video). In this case, the problem described in Equation 22 finds the appropriate non-linear transformation that maps the input features to the *aligned* temporal labels. Furthermore, another example of utilising the proposed DDATW formulation lies in settings where the alignment of multiple sequences is required while at the same time, discrete temporal labels are readily available. In this scenario, we can obtain the appropriate non-linear, discriminative transformation during training, by utilising the provided labels[5]. Given an out-of-sample sequence during testing, we can then extract the non-linear transformations (learned while utilising labels available during training by solving Equation 22) and subsequently estimate the optimal time-warpings ($\boldsymbol{\Delta}_i$) that align the out-of-sample sequences to the learnt discriminative subspace.

5 Experiments

In order to assess the performance of DCTW, we perform detailed experiments against both linear and non-linear state-of-the-art temporal alignment algorithms. In more detail we compare against:

State of the art methods for time warping without a feature extraction step:

- Dynamic Time Warping (DTW) [33] which finds the optimal alignment path given that the sequences reside in the same manifold (as explained in subsection 2.2).
- Iterative Motion Warping (IMW) [21] alternates between time warping and spatial transformation to align two sequences.

State-of-the art methods with a linear feature extractor:

- Canonical Time Warping (CTW) [43] as posed in section subsection 2.2, CTW finds the optimal reduced dimensionality subspace such that the sequences are maximally linearly correlated.
- Generalized Time Warping (GTW) [44] which uses a combination of CTW and a Gauss-Newton temporal warping method that parametrises the warping path as a combination of monotonic functions.

State-of-the-art methods with non-linear feature extraction process.

- Manifold Time Warping [40] that employs a variation of Laplacian Eigenmaps to non-linearly transform the original sequences.

We evaluate the aforementioned techniques on four different real-world datasets, namely *(i)* the Weizmann database subsection 5.2, where multiple feature sets are aligned , *(ii)* the MMI Facial Expression database subsection 5.3, where we apply DCTW on the alignment of facial Action Units, *(iii)* the XRMB database subsection 5.4 where we align acoustic and articulatory recordings, and finally *(iv)* the CUAVE database subsection 5.5, where we align visual and auditory utterances.

Evaluation For all experiments, unless stated otherwise, we assess the performance of DCTW utilising the the alignment error introduced in [44]. Assuming we have m sequences, each algorithm infers a set of warping paths $\mathbf{P}_{\text{alg}} = \left[\mathbf{p}_1^{\text{alg}}, \mathbf{p}_2^{\text{alg}}, \ldots, \mathbf{p}_m^{\text{alg}}\right]$, where $\mathbf{p}_i \in \left\{x \in \mathbb{N}^{l_{\text{alg}}} | 1 \leq x \leq n_m\right\}$ is the alignment path for the ith sequence with a length l_{alg}. The error is then defined as

$$\text{Err} = \frac{\text{dist}(\mathbf{P}^{\text{alg}}, \mathbf{P}^{\text{ground}}) + \text{dist}(\mathbf{P}^{\text{ground}}, \mathbf{P}^{\text{alg}})}{l_{\text{alg}} + l_{\text{ground}}},$$

$$\text{dist}\left(\mathbf{P}^1, \mathbf{P}^2\right) = \sum_{i=1}^{l_1} \min_{j=1}^{l_2} \left\|\mathbf{p}_{(i)}^1 - \mathbf{p}_{(j)}^2\right\|_2.$$

5.1 Experimental Setup

In each experiment, we perform unsupervised pretraining of the deep architecture for each of the available sequences in order to speed up the convergence of the optimisation procedure. In particular, we initialise the parameters of each of the layers using a denoising autoencoder [39]. We utilise full-batch optimisation with AdaGrad [12] for training, although similar results are obtained by utilising mini-batch stochastic gradient descent optimisation with a large mini-batch size. In contrast to [2], we utilise a leaky rectified linear unit with $a = 0.03$ (LReLU) [26], where $f(x) = \max(ax, x)$ and a is a small positive value. In our experiments, this function converged faster and produced better results than the suggested modified cube-root sigmoid activation function. For all the experiments (excluding subsection 5.2 where a smaller network was sufficient) we utilised a fixed three layer 200–100–100 fully connected topology, thus reducing the number of free hyperparameters of the architecture.

5. If the labels for any subset $\mathcal{K}$ of available sequences are considered to be aligned with the corresponding features, then we can simply set $\forall i.\boldsymbol{\Delta}_i = \mathbf{I}$ where $i \in \mathcal{K}$.

This both facilitates the straight-forward reproducibility of experimental results, as well as helps towards avoiding overfitting (particularly since training is unsupervised).

5.2 Real Data I: Alignment of Human Actions under Multiple Feature Sets

In this experiment, we utilise the Weizmann database [17], containing videos of nine subjects performing one of ten actions (e.g., walking). We adopt the experimental protocol described in [44], where 3 different shape features are computed for each sequence, namely *(1)* a binary mask, *(2)* Euclidean distance transform [28], and *(3)* the solution of the Poisson equation [16], [44]. Subsequently, we reduce the dimensionality of the frames to 70–by–35 pixels, while we keep the top 123 principle components. For all algorithms, the same hyperparameters as [44] are used. Following [43], [44], 90% of the total correlation is kept, while we used a topology of two layers carrying 50 neurons each. Triplets of videos where subjects are performing the same action where selected, and each alignment algorithm was evaluated on aligning the three videos based on the features described above.

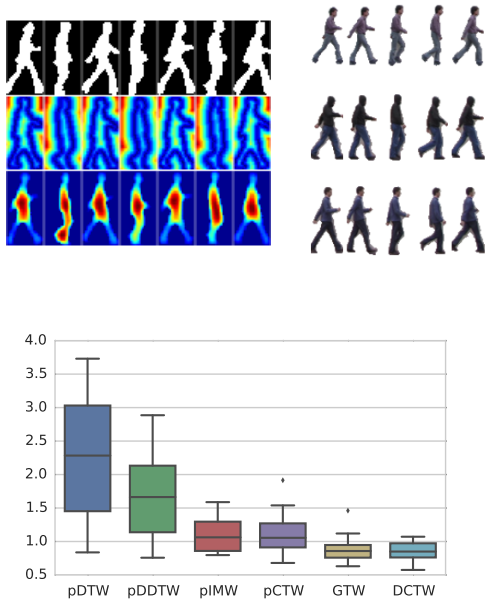

Fig. 3. Aligning sequences of subjects performing similar actions from the Weizmann database. (left) the three computed features for each of the sequences (1) binary (2) euclidean (3) poisson solution. (middle) The aligned sequences using DCTW. (right) Alignment errors for each of the six techniques.

The ground truth of the data was approximated by running DTW on the binary mask images. Thus, the reasoning behind this experiment is to evaluate whether the methods manage to find a correlation between the three computed features, in which case they would find the alignment path produced by DTW.

In Figure 3 we show the alignment error for ten randomly generated sets of videos. As DTW, DDATW, IMW, and CTW are only formulated for performing alignment between two sequences we use their multi-sequence extension as formulated in [45] and we use the prefix p to denote the multisequence variant.

We observe that DTW and DDTW fail to align the videos correctly, while CTW, GTW, and DCTW perform quite better. This can be justified by considering that DTW and DDTW are applied directly on the observation space, while CTW, GTW and DCTW infer a common subspace of the three input sequences. The best performing methods are clearly GTW and DCTW.

5.3 Real Data II: Alignment of Facial Action Units

Next, we evaluate the performance of DCTW on the task of temporal alignment of facial expressions. We utilise the MMI Facial Expression Dataset [31] which contains more than 2900 videos of 75 different subjects, each performing a particular combination of Action Units (*i.e.*, facial muscle activations). We have selected a subset of the original dataset which contains videos of subjects which manifest the same action unit (namely, AU12 which corresponds to a smile), and for which we have ground truth annotations. We preprocessed all the images by converting to greyscale and utilised an off-the-shelf face detector along with a face alignment procedure [23] in order to crop a bounding box around the face of each subject. Subsequently, we reduce the dimensionality of the feature space to 400 components using whitening PCA, preserving 99% of the energy. We clarify that the annotations are given for each frame, and describe the temporal phase of the particular AU at that frame. Four possible temporal phases of facial action units are defined: *neutral* when the corresponding facial muscles are inactive, *onset* where the muscle is activated, *apex* when facial muscle intensity reaches its peak, and *offset* when the facial muscle begins to relax, moving towards the neutral state. Utilising *raw* pixels, the goal of this experiment lies in temporally aligning each pair of videos. In the context of this experiment, this means that the subjects in both videos exhibit the same temporal phase at the same time. E.g., for smiles, when subject 1 in video 1 reaches the apex of the smile, the subject in video 2 does so as well. In order to quantitatively evaluate the results, we utilise the ratio of correctly aligned frames within each temporal phase to the total duration of the temporal phase across the aligned videos. This can be formulated as $\frac{|\Phi_1 \cap \Phi_2|}{|\Phi_1 \cup \Phi_2|}$, where $\Phi_{1,2}$ is the set of aligned frame indices after warping the initial vector of annotations using the alignment matrices $\mathbf{\Delta}_i$ found via a temporal warping technique.

Results are presented in Figure 4, where we illustrate the alignment error on 45 pairs of videos across all methods and action unit temporal phases. Clearly, DTW overperforms MW, while CCA based methods such as CTW and GTW perform better than DTW. It can be seen that the best performance in all cases is obtained by DCTW, and using a t-test with the next best method we find that the result is statistically significant ($p < 0.05$). This can be justified by the fact that the non-linear hierarchical structure of DCTW facilitates the modelling of the complex dynamics straight from the low-level pixel intensities.

Furthermore, in Figure 5 we illustrate the alignment results from a pair of videos of the dataset. The first row depicts the first sequence in the experiment, where for each

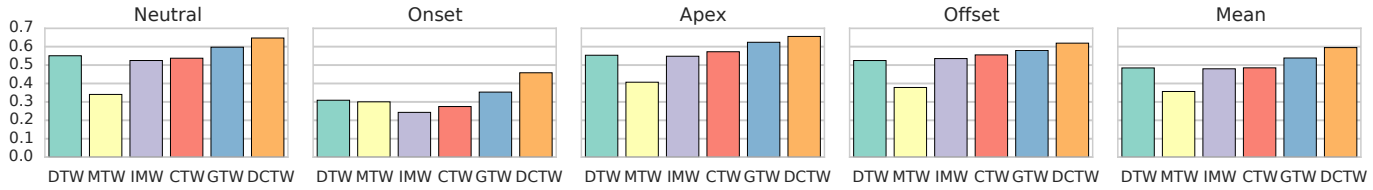

Fig. 4. Temporal phase detection accuracy as defined by the ratio of correctly aligned frames with respect to the total duration for each temporal phase – the higher the better.

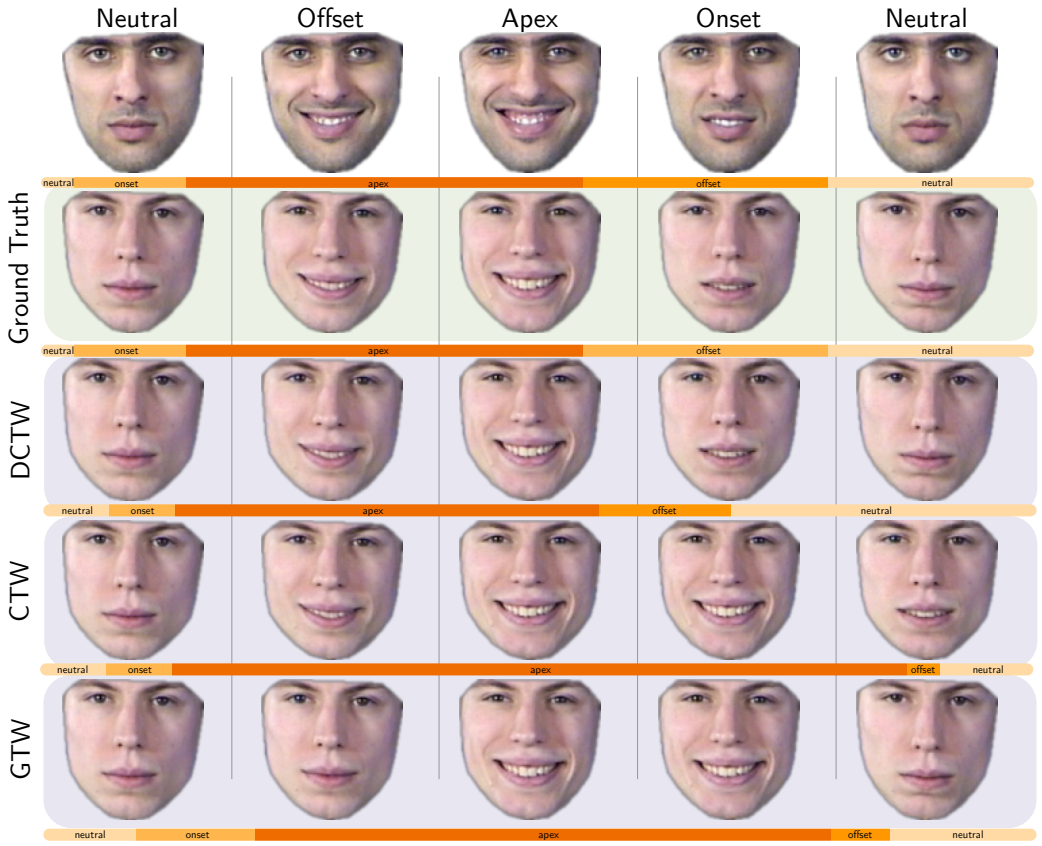

Fig. 5. Facial expression alignment of videos S002–005 and S014–009 from MMI dataset (subsection 5.3). Depicted frames for each temporal phase with duration $[t_s, t_e]$ correspond to the middle of each of the temporal phase, $t_c = \lceil \frac{t_s+t_e}{2} \rceil$. We also plot the temporal phases (● neutral, ● onset, ● apex, and ● offset) corresponding to (i) the ground truth alignment and (ii) compared methods (DCTW, CTW and GTW). Note that the entire video is included in our supplementary material.

temporal phase with duration $[t_s, t_e]$ we plot the frame $t_c = \lceil \frac{t_s+t_e}{2} \rceil$. The second row illustrates the ground truth of the second video, while the following rows compare the alignment paths obtained by DCTW, CTW and GTW respectively. By observing the corresponding images as well as the temporal phase overlap, it is clear that DCTW achieves the best alignment. At last we repeat the experiment using a convolutional network topology which operates directly on the raw image pixel intensities. We opted for a simple architecture similar to LeNet [35], consisting of 2 convolutional layers of 32 filters each (kernel size 3) followed by a 2x2 max-pooling operation and finally a linear projection to 10 dimensions. Although this architecture attains the same performance in terms of accuracy, we found that (i) the optimisation converged quicker, and (ii) we obtained interpretable features which show the inner-workings of the network shown in Figure 6.

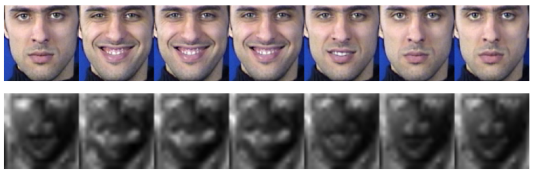

Fig. 6. Depicted are the last convolutional features (bottom row) using a 3-layer architecture showing frames from a video (top row) containing AU12 (Lip Corner Puller). The features seem to activate on the presence of smile and squinting of the eyes.

5.4 Real Data III: Alignment of Acoustic and Articulatory Recordings

The third set of experiments involves aligning simultaneous acoustic and articulatory recordings from the Wisconsin X-ray Microbeam Database (XRMB) [42]. The articulatory data

consist of horizontal and vertical displacements of eight pellets on the speaker's lips, tongue, and jaws, yielding a 16-dimensional vector at each time point. We utilise the features provided by [2]. The baseline acoustic features consist of standard 13-dimensional mel-frequency cepstral coefficients (MFCCs) [8] and their first and second derivatives computed every 10ms over a 25ms window. For the articulatory measurements to match the MFCC rate, we concatenate them over a 7-frame window, thus obtaining $\mathbf{X}_{\text{art}} \in \mathbb{R}^{273}$ and $\mathbf{X}_{\text{MFCC}} \in \mathbb{R}^{112}$.

As the two views were recorded simultaneously and then manually synchronised [42], we use this correspondence as the ground truth and then we produce a synthetic misalignment to the sequences, producing 10 sequences of 5000 samples. We warp the auditory features using the alignment path produced by $\mathcal{P}_{\text{mis}}(i) = i^{1.1} l_{\text{MFCC}}^{0.1}$ for $1 \leq i \leq l_{\text{MFCC}}$ where l_{MFCC} is the number of MFCC samples.

Results are presented in Table 1. Note that DCTW outperforms compared methods by a much larger margin than other experiments here. Nevertheless, this is quite expected: the features for this experiment are highly heterogeneous and e.g., in case of MFCCs, non-linear. The multi-layered non-linear transformations applied by DCTW are indeed much more suitable for modelling the mapping between such varying feature sets.

DTW	MTW	IMW
63.52 ± 27.06	94.42 ± 13.20	83.23 ± 0.11
CTW	GTW	DCTW
58.92 ± 28.8	64.06 ± 5.01	7.19 ± 1.79

TABLE 1
Alignment errors obtained on the Wisconsin X-ray Microbeam Database.

5.5 Real Data IV: Alignment of Audio and Visual Streams

In arguably, our most challenging experimental setting, we aim to align the subject's visual and auditory utterances. To this end, we use the CUAVE [32] database which contains 36 videos of individuals pronouncing the digits 0 to 9. In particular, we use the portion of videos containing only frontal facing speakers pronouncing each digit five times, and use the same approach as in subsection 5.4 in order to introduce misalignments between the audio and video streams. In order to learn the hyperparameters of all employed alignment techniques, we leave out 6 videos.

Regarding pre-processing, from each video frame we extract the region-of-interest (ROI) containing the mouth of the subject using the landmarks produced via [23]. Each ROI was then resized to 60 x 80 pixels, while we keep the top 100 principal components of the original signal. Subsequently, we utilise temporal derivatives over the reduced vector space. Regarding the audio signal, we compute the Mel-frequency cepstral coefficients (MFCC) features using a 25ms window adopting a step size of 10ms between successive windows. Finally, we compute the temporal derivatives over the acoustic features (and video frames). To match the video frame rate, 3 continuous audio frames are concatenated in a vector. The results show that DCTW

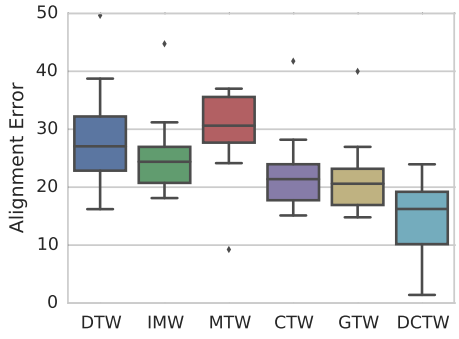

Fig. 7. Alignment errors on the task of audio-visual temporal alignment. Note that videos better illustrating the results are included in our supplementary material.

outperforms the rest of the temporal alignment methods by a large margin. Again, the justification is similar to subsection 5.4: the highly heterogeneous nature of the acoustic and video features highlights the significance of deep non-linear architectures for the task-at-hand. It should be noted that the best results obtained for GTW utilise a combination of hyperbolic and polynomial basis, which biases the results in favour of GTW due to the misalignment we introduce. Still, it is clear that DCTW obtains much better results in terms of alignment error.

5.6 Deep Discriminant Analysis with Time Warpings (DDATW)

We perform two additional experiments in order to evaluate the proposed Deep Discriminant Analysis (Section 4), where in this case our data will also consist of a set of labels corresponding to the samples at-hand. In our first experiment, we utilise set of videos described in subsection 5.3 from the MMI database, that display facial expressions. In more detail, we exploit the fact that the action units have been labelled with regards to the temporal phases of facial behaviour (similarly to the setting for the experiment described in subsection 5.3.) Since each frame of the video has been assigned to a temporal phase, we utilise these labels in order to evaluate the proposed DDA. In particular, during training we utilise the available labels in order to learn the discriminant transformation. Subsequently, the learnt transformation can be applied to testing data in order to predict the labels. An example of the temporal segment annotations and the corresponding prediction for AU12 (Lip Corner Puller) can be found in Figure 2.

For our second experiment, we utilise the temporal labels available for the CUAVE dataset (as described in subsection 5.5). Since in each video a subject is uttering the digits 1 to 10, each framed is labelled with respect to whether the subject is uttering a digit or not. If the subject is uttering a digit, then the corresponding class corresponds to the particular digit being uttered. If not, then the frame is classified separately. This leads to 11 classes in total. For this experiment, we utilise half the data for training/validation and the other half for testing. The results are summarised in Table 2, where we compare between the unsupervised CTW and DCTW, as well as the proposed Deep Discriminant Analysis with Time Warpings (DDATW), as well as the

linear version of DDATW, which we term DATW. Note that by introducing supervision, we are able to further improve results.

TABLE 2
Classification accuracy using the available temporal phase labels for MMI (3 labels) and the digit annotations for CUAVE (11 labels).

	CTW	DATW	DCTW	DDATW
MMI	49.2	53.5	59.1	65.1
CUAVE	35.7	43.6	68.7	83.7

6 Computational details and discussion

The computational complexity of aligning a set of m sequences each of length T_i is $\mathcal{O}(\sum_{i,j}^{m} T_i T_j + \sum_{i=1}^{m} d_i^3)$ per iteration of the algorithm, which is the complexity of DTW plus the cost of the SVD in the computation of the derivatives in Equation 15. As the SVD is performed on the last layer of the network, which is of reduced dimensionality ($d_i = 100$ units in our case) it is relatively cheap. In contrast other non-linear warping algorithms [40] require an expensive k-nearest neighbour search accompanied by an eigendecomposition step or, in the case of CTW [43], an eigendecomposition of the original covariance matrices which becomes much more expensive when dealing with data of high dimensionality. Nevertheless, the proposed algorithm may require to perform more iterations in order to converge than CTW. In particular, DCTW needed around 5 minutes to converge in our second experiment subsection 5.3 of aligning facial action units, while for the same experiment CTW required around 1 minute. A way to expedite the procedure is to apply linear approximations of DTW such as [34], [44] or optimise the alignment paths only on a subset of iterations (this is an interesting line of further research on the topic).

Finally, it is worthwhile to mention that although in this work we explored simple network topologies, our cost function can be optimised regardless of the number of layers or neuron type (e.g., convolutional). Finally we also note that DCTW is agnostic to the use of the method for temporally warping the sequences and other relaxed variants of DTW might be employed in practise when there is a large number of observations in each sequence as for example Fast DTW [34] or GTW [44] as long as it conforms to the alignment constrains, i.e., it always minimises the objective function.

7 Conclusions

In this paper, we study the problem of temporal alignment of multiple sequences. To the best of our knowledge, we propose the first temporal alignment method based on deep architectures, which we dub Deep Canonical Time Warping (DCTW). DCTW discovers a hierarchical non-linear feature transformation for multiple sequences, where (i) all transformed features are temporally aligned, and (ii) are maximally correlated. Furthermore, we consider the setting where temporal labels are provided for the data-at-hand. By modifying the objective function for the proposed method, we are able to provide discriminant feature mappings that may be more suitable for classification tasks. Finally, by means of various experiments on several datasets, we highlight the significance of the proposed methods on various applications, as the proposed method outperforms compared state-of-the-art methods.

Acknowledgments

George Trigeorgis is a recipient of the fellowship of the Department of Computing, Imperial College London, and this work was partially funded by it. The work of Stefanos Zafeiriou was partially funded by the EPSRC project EP/J017787/1 (4D-FAB), as well as by the FiDiPro program of Tekes (project number: 1849/31/2015). The work of Björn W. Schuller was partially funded by the European Community's Horizon 2020 Framework Programme under grant agreement No. 645378 (ARIA-VALUSPA). The responsibility lies with the authors.

George Trigeorgis has received an MEng in Artificial Intelligence in 2013 from the Department of Computing, Imperial College London, where he is currently completing his Ph.D. studies. He was a recipient of the prestigious Google Ph.D. Fellowship in Machine Perception for 2017. He has regularly published in several prestigious conferences in his field including ICML, NIPS, and IEEE CVPR, while he is also a reviewer in IEEE T-PAMI, IEEE CVPR/ICCV/FG.

Mihalis A. Nicolaou is a Lecturer at the Department of Computing at Goldsmiths, University of London and an Honorary Research Fellow with the Department of Computing at Imperial College London. Mihalis obtained his PhD from the same department at Imperial, while he completed his undergraduate studies at the Department of Informatics and Telecommunications at the University of Athens, Greece. Mihalis' research interests span the areas of machine learning, computer vision and affective computing. He has been the recipient of several awards and scholarships for his research, including a Best Paper Award at IEEE FG, while publishing extensively in related prestigious venues. Mihalis served as a Guest Associate Editor for the IEEE Transactions on Affective Computing and is a member of the IEEE.

Stefanos Zafeiriou is currently a Senior Lecturer in Pattern Recognition/Statistical Machine Learning for Computer Vision with the Department of Computing, Imperial College London, London, U.K, and a Distinguishing Research Fellow with University of Oulu under Finish Distinguishing Professor Programme. He was a recipient of the Prestigious Junior Research Fellowships from Imperial College London in 2011 to start his own independent research group. He was the recipient of the President's Medal for Excellence in Research Supervision for 2016. He has received various awards during his doctoral and post-doctoral studies. He currently serves as an Associate Editor of the IEEE Transactions on Cybernetics the Image and Vision Computing Journal. He has been a Guest Editor of over six journal special issues and co-organised over nine workshops/special sessions on face analysis topics in top venues, such as CVPR/FG/ICCV/ECCV (including two very successfully challenges run in ICCV13 and ICCV15 on facial landmark localisation/tracking). He has more than 2800 citations to his work, h-index 27. He is the General Chair of BMVC 2017.

Björn W. Schuller received his diploma, doctoral degree, habilitation, and Adjunct Teaching Professorship all in EE/IT from TUM in Munich/Germany. At present, he is a Reader (Associate Professor) in Machine Learning at Imperial College London/UK, Full Professor and Chair of Complex & Intelligent Systems at the University of Passau/Germany, and the co-founding CEO of audEERING UG. Previously, he headed the Machine Intelligence and Signal Processing Group at TUM from 2006 to 2014. In 2013 he was also invited as a permanent Visiting Professor at the Harbin Institute of Technology/P.R. China and the University of Geneva/Switzerland. In 2012 he was with Joanneum Research in Graz/Austria remaining an expert consultant. In 2011 he was guest lecturer in Ancona/Italy and visiting researcher in the Machine Learning Research Group of NICTA in Sydney/Australia. From 2009 to 2010 he was with the CNRS-LIMSI in Orsay/France, and a visiting scientist at Imperial College. He co-authored 500+ technical contributions (10,000+ citations, h-index = 49) in the field.

REFERENCES

[1] J. Aach and G. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17:495–508, 2001.
[2] G. Andrew et al. Deep Canonical Correlation Analysis. In *ICML*, volume 28, 2013.
[3] F. Bach. Consistency of trace norm minimization. *JMLR*, 9:1019–1048, 2008.
[4] F. Bach and M. Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.
[5] A. Bruderlin and L. Williams. Motion signal processing. In *SIGGRAPH*, pages 97–104, 1995.
[6] Y. Caspi and M. Irani. Aligning non-overlapping sequences. *IJCV*, 48:39–51, 2002.
[7] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE CVPR*. IEEE, 2005.
[8] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE TASSP*, 28, 1980.
[9] F. De La Torre. A least-squares framework for component analysis. *IEEE TPAMI*, 34:1041–1055, 2012.
[10] F. De La Torre. A least-squares framework for component analysis. *IEEE TPAMI*, 34(6):1041–1055, 2012.
[11] M. Dorfer, R. Kelz, and G. Widmer. Deep linear discriminant analysis. *arXiv preprint arXiv:1511.04707*, 2015.
[12] J. Duchi et al. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *JMLR*, 12:2121–2159, 2011.
[13] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 2013.
[14] R. Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE CVPR*, pages 580–587. IEEE, 2014.
[15] D. Gong and G. Medioni. Dynamic Manifold Warping for view invariant action recognition. In *IEEE CVPR*, pages 571–578, 2011.
[16] L. Gorelick et al. Shape representation and classification using the poisson equation. *IEEE TPAMI*, 28:1991–2004, 2006.
[17] L. Gorelick et al. Actions as space-time shapes. *IEEE TPAMI*, 29:2247–2253, 2007.
[18] M. Hasan. On multi-set canonical correlation analysis. In *IJCNN*, pages 1128–1133. IEEE, 2009.
[19] G. Hinton et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Sig. Prog. Mag.*, 29(6):82–97, 2012.
[20] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.
[21] E. Hsu, K. Pulli, and J. Popović. Style translation for human motion. In *SIGGRAPH*, volume 24, page 1082, 2005.
[22] B.-H. F. Juang. On the hidden markov model and dynamic time warping for speech recognitiona unified view. *AT&T Bell Laboratories Technical Journal*, 63(7):1213–1243, 1984.
[23] V. Kazemi and S. Josephine. One Millisecond Face Alignment with an Ensemble of Regression Trees. In *IEEE CVPR*, 2014.
[24] Y. Kim, H. Lee, and E. M. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *IEEE ICASSP*, pages 3687–3691. IEEE, 2013.
[25] A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.
[26] A. Maas et al. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, volume 30, 2013.
[27] K. Mardia et al. *Multivariate analysis*. Academic press, 1979.
[28] C. Maurer and V. Raghavan. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. *IEEE TPAMI*, 25:265–270, 2003.
[29] J. Ngiam, A. Khosla, and M. Kim. Multimodal deep learning. *ICML*, 2011.
[30] A. A. Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE TIP*, 11(3):293–305, 2002.
[31] M. Pantic et al. Web-based database for facial expression analysis. In *ICME*, volume 2005, pages 317–321, 2005.
[32] E. Patterson et al. CUAVE: A new audio-visual database for multimodal human-computer interface research. *ICASSP*, 2:II–2017–II–2020, 2002.
[33] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*, volume 103. 1993.
[34] S. Salvador and P. Chan. Fastdtw: Toward accurate dynamic time warping in linear time and space. In *KDD-Workshop*, 2004.
[35] B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
[36] S. Shariat and V. Pavlovic. Isotonic CCA for sequence alignment and activity recognition. In *ICCV*, pages 2572–2578, 2011.

[37] Y. Taigman et al. Deepface: Closing the gap to human-level performance in face verification. In *IEEE CVPR*, pages 1701–1708. IEEE, 2014.
[38] G. Trigeorgis, M. Nicolaou, S. Zafeiriou, and B. W. Schuller. Deep Canonical Time Warping. In *CVPR*, 2016.
[39] P. Vincent et al. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *JMLR*, 11:3371–3408, 2010.
[40] H. Vu et al. Manifold Warping: Manifold Alignment over Time. In *AAAI*, 2012.
[41] A. Wang, J. Lu, G. Wang, J. Cai, and T.-J. Cham. Multi-modal unsupervised feature learning for RGB-D scene labeling. In *ECCV*, pages 453–467. Springer, 2014.
[42] J. Westbury et al. X-ray microbeam speech production database. *JASA*, 88(S1):S56—-S56, 1990.
[43] F. Zhou and F. De La Torre. Canonical time warping for alignment of human behavior. *NIPS*, 2009.
[44] F. Zhou and F. De La Torre. Generalized time warping for multi-modal alignment of human motion. In *IEEE CVPR*, pages 1282–1289, 2012.
[45] F. Zhou and F. De La Torre. Generalized Canonical Time Warping. *IEEE TPAMI*, 2015.