Deep Canonical Time Warping

George Trigeorgis[1] Mihalis A. Nicolaou[2] Stefanos Zafeiriou[1,3]
Björn W. Schuller[1]
[1]Imperial College London, UK [2]Goldsmiths, University of London,U K
[3] Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[1]{g.trigeorgis, s.zafeiriou, bjoern.schuller}@imperial.ac.uk, [2]m.nicolaou@gold.ac.uk

Abstract

Machine learning algorithms for the analysis of time-series often depend on the assumption that the 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 the 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 the temporal alignment of time-series are applied directly on the observation space, or utilise simple linear projections. Thus, they fail to capture complex, hierarchical non-linear representations which may prove to be beneficial towards the task of temporal alignment, particularly when dealing with multi-modal data (e.g., aligning visual and acoustic information). To this end, we present the Deep Canonical Time Warping (DCTW), a method which automatically learns complex non-linear representations of multiple time-series, generated such that (i) they are highly correlated, and (ii) temporally in alignment. By means of experiments on four real datasets, we show that the representations learnt via the proposed DCTW significantly outperform state-of-the-art methods in temporal alignment, elegantly handling scenarios with highly heterogeneous features, such as the temporal alignment of acoustic and visual features.

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 [29], computer vision [6], graphics [5] and bio-informatics [1]. Example applications range from the temporal alignment of facial expressions and motion capture data [37, 38], to the alignment for human action recognition [34], and speech [18].

The most prominent temporal alignment method is Dynamic Time Warping (DTW) [29], 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 [37] 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 [34, 11] while in [31, 38] constraints are introduced in order to guarantee monotonicity and adaptively constrain the temporal warping. It should be noted that in [38], a multi-set variant of CCA is utilised [14] 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 [16], feature extraction for object recognition and detection [21, 10], feature extraction for face recognition [32], acoustic modelling in speech recognition [15], as well as for extracting non-linear correlated features [2].

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 [20, 25], and on robotics applications with visual and depth data [35].

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, which in addition offers very good scaling when dealing with large amounts of data. In more detail, our work carries the following contributions: (i) we extend DTW-based temporal alignment methods to handle heterogeneous collections of features which 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 finally (iii) we evaluate the proposed DCTW 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. In Sec. 3 we refer to related work on temporal alignment and canonical correlation analysis. In Sec. 4, we describe the proposed DCTW. We provide experiments on four real datasets in Sec. 5, while we conclude the paper in Sec. 7.

2. Notation

Throughout the paper, matrices are denoted by uppercase boldface letters (e.g., $\mathbf{X}, \mathbf{Y}$), vectors are denoted by lowercase boldface letters (e.g., $\mathbf{x}, \mathbf{y}$), and scalars appear as either uppercase or lowercase letters.

3. Related Work

3.1. 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

$$\begin{aligned}\boldsymbol{\rho} &= \frac{\mathbb{E}[\mathbf{Y}_1\mathbf{Y}_2^\top]}{\sqrt{\mathbb{E}[\mathbf{Y}_1\mathbf{Y}_1^\top\mathbf{Y}_2\mathbf{Y}_2^\top]}} && (1)\\ &= \frac{\mathbf{W}_1^\top\mathbb{E}[\mathbf{X}_1\mathbf{X}_2^\top]\mathbf{W}_2}{\sqrt{\mathbf{W}_1^\top\mathbb{E}[\mathbf{X}_1\mathbf{X}_1^\top]\mathbf{W}_1\mathbf{W}_2^\top\mathbb{E}[\mathbf{X}_2\mathbf{X}_2^\top]\mathbf{W}_2}} && (2)\\ &= \frac{\mathbf{W}_1^\top\boldsymbol{\Sigma}_{12}\mathbf{W}_2}{\sqrt{\mathbf{W}_1^\top\boldsymbol{\Sigma}_{11}\mathbf{W}_1\mathbf{W}_2^\top\boldsymbol{\Sigma}_{22}\mathbf{W}_2}} && (3)\end{aligned}$$

where $\boldsymbol{\Sigma}_{ij}$ denotes the empirical covariance between data matrices $\mathbf{X}_i$ and $\mathbf{X}_j$[1]. There are multiple equivalent optimisation problems for discovering the optimal loadings $\mathbf{W}_i$ which maximise Eq. 3 [8]. For instance, CCA can be formulated as a least-squares problem,

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

and equivalently as a trace optimisation problem

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

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

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{W}_1 = \boldsymbol{\Sigma}_{11}\mathbf{W}_1\boldsymbol{\Lambda}. \qquad (6)$$

Note that an equivalent solution is obtained by resorting to Singular Value Decomposition (SVD) on the matrix $\mathbf{K} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$[23, 4]. The optimal objective value of Eq. 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 Eq. 6 and keeping the top d eigenvectors corresponding to the largest eigenvalues.

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

[1] Note that we assume zero-mean data to avoid cluttering the notation.

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

$$\arg\max_{\theta_1,\theta_2} \operatorname{corr}(f_1(\mathbf{X}_1;\theta_1), f_2(\mathbf{X}_2;\theta_2)). \tag{7}$$

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 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.*, $\tilde{\mathbf{\Sigma}}_{ij} = \frac{1}{T-1}\tilde{\mathbf{X}}_i(\mathbf{I} - \frac{1}{T}\mathbf{1}\mathbf{1}^\top)\tilde{\mathbf{X}}_i^\top$. As described above for classical CCA (Eq. 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 $\mathbf{K}$, $\|\mathbf{K}\|_* = \operatorname{trace}(\sqrt{\mathbf{K}\mathbf{K}^\top})$. Problem 7 now becomes

$$\arg\max_{\theta_1,\theta_2} \|\mathbf{K}\|_* \,. \tag{8}$$

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

3.2. 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 Eq. 9,

$$\begin{aligned} &\arg\min_{\mathbf{\Delta}_1,\mathbf{\Delta}_2} \|\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{X}_2\mathbf{\Delta}_2\|_F^2 \\ &\text{subject to:} \quad \mathbf{\Delta}_1 \in \{0,1\}^{T_1\times T}, \\ &\qquad\qquad\quad \mathbf{\Delta}_2 \in \{0,1\}^{T_2\times T}, \end{aligned} \tag{9}$$

where $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$ are binary selection matrices [37] 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 Eq. 9) in $\mathcal{O}(T_1T_2)$. Finally, the DTW solution satisfies the boundary, continuity, and monotonicity constraints [29].

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) [37] elegantly combines the least-squares formulations of DTW (Eq. 9) and CCA (Eq. 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

$$\begin{aligned} &\arg\min_{\mathbf{W}_1,\mathbf{W}_2,\mathbf{\Delta}_1,\mathbf{\Delta}_2} \|\mathbf{W}_1^\top\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{W}_2^\top\mathbf{X}_2\mathbf{\Delta}_2\|_F^2 \\ &\text{subject to:} \quad \mathbf{W}_1^\top\mathbf{X}_1\mathbf{\Delta}_1\mathbf{\Delta}_1^\top\mathbf{X}_1^\top\mathbf{W}_1 = \mathbf{I}, \\ &\qquad \mathbf{W}_2^\top\mathbf{X}_2\mathbf{\Delta}_2\mathbf{\Delta}_2^\top\mathbf{X}_2^\top\mathbf{W}_2 = \mathbf{I}, \\ &\qquad \mathbf{W}_1^\top\mathbf{X}_1\mathbf{\Delta}_1\mathbf{\Delta}_2^\top\mathbf{X}_2^\top\mathbf{W}_2 = \mathbf{D}, \\ &\qquad \mathbf{X}_1\mathbf{\Delta}_1\mathbf{1} = \mathbf{X}_2\mathbf{\Delta}_2\mathbf{1} = \mathbf{0} \\ &\qquad \mathbf{\Delta}_1 \in \{0,1\}^{T_1\times T}, \mathbf{\Delta}_2 \in \{0,1\}^{T_2\times T} \end{aligned} \tag{10}$$

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 Eq. 10, mostly inherited by CCA, deem the CTW solution translation, rotation, and scaling invariant. A solution is then subsequently obtained by alternating between solving CCA (by fixing $\mathbf{X}_i\mathbf{\Delta}_i$) and DTW (by fixing $\mathbf{W}_i^\top\mathbf{X}_i$).

4. Deep Canonical Time Warping

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 $\mathbf{X}_i$. We propose to optimise the following objective,

$$\begin{aligned} &\arg\min_{\theta_1,\theta_2,\mathbf{\Delta}_1,\mathbf{\Delta}_2} \|f_1(\mathbf{X}_1;\theta_1)\mathbf{\Delta}_1 - f_2(\mathbf{X}_2;\theta_2)\mathbf{\Delta}_2\|_F^2 \\ &\text{subject to:} \quad f_1(\mathbf{X}_1;\theta_1)\mathbf{\Delta}_1\mathbf{\Delta}_1^\top f_1(\mathbf{X}_1;\theta_1)^\top = \mathbf{I}, \\ &\qquad f_2(\mathbf{X}_2;\theta_2)\mathbf{\Delta}_2\mathbf{\Delta}_2^\top f_2(\mathbf{X}_2;\theta_2)^\top = \mathbf{I}, \\ &\qquad f_1(\mathbf{X}_1;\theta_1)\mathbf{\Delta}_1\mathbf{\Delta}_2^\top f_2(\mathbf{X}_2;\theta_2) = \mathbf{D}, \\ &\qquad f_1(\mathbf{X}_1;\theta_1)\mathbf{\Delta}_1\mathbf{1} = f_2(\mathbf{X}_2;\theta_2)\mathbf{\Delta}_2\mathbf{1} = \mathbf{0}, \\ &\qquad \mathbf{\Delta}_1 \in \{0,1\}^{T_1\times T}, \mathbf{\Delta}_2 \in \{0,1\}^{T_2\times T} \end{aligned} \tag{11}$$

where as defined for Eq. 10, $\mathbf{D}$ is a diagonal matrix and $\mathbf{1}$ is an appropriate dimensionality vector of all 1's. 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 find-

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

ing the optimal parameters of each network by utilising backpropagation. Having discovered the warping matrices $\boldsymbol{\Delta}_i$, 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 $\hat{\boldsymbol{\Sigma}}_{ij} = \frac{1}{T-1} f_i(\mathbf{X}_i;\theta_i)\boldsymbol{\Delta}_i\mathbf{C}_T\boldsymbol{\Delta}_j^\top f_j(\mathbf{X}_j;\theta_j)^\top$, where $\mathbf{C}_T$ is the centring matrix, $\mathbf{C}_T = \mathbf{I} - \frac{1}{T}\mathbf{1}\mathbf{1}^\top$. By defining $\mathbf{K}_{\mathcal{DCTW}} = \hat{\boldsymbol{\Sigma}}_{11}^{-1/2}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{\Sigma}}_{22}^{-1/2}$, we now have that

$$\operatorname{corr}(f_1(\mathbf{X}_1;\theta_1)\boldsymbol{\Delta}_1, f_2(\mathbf{X}_2;\theta_2)\boldsymbol{\Delta}_2) = \|\mathbf{K}_{\mathcal{DCTW}}\|_*. \tag{12}$$

We optimise this quantity in a gradient-ascent fashion by utilising the subgradient of Eq. 12 [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 $\mathbf{USV}^\top = \mathbf{K}_{\mathcal{DCTW}}$ is the singular value decomposition of $\mathbf{K}_{\mathcal{DCTW}}$, then the subgradient for the last layer is defined as

$$\begin{aligned}
\mathbf{F}^{(\text{pos})} &= \hat{\boldsymbol{\Sigma}}_{11}^{-1/2}\mathbf{U}\mathbf{V}^\top\hat{\boldsymbol{\Sigma}}_{22}^{-1/2}\mathbf{Y}_2\boldsymbol{\Delta}_2\mathbf{C}_T\\
\mathbf{F}^{(\text{neg})} &= \hat{\boldsymbol{\Sigma}}_{11}^{-1/2}\mathbf{U}\mathbf{S}\mathbf{U}^\top\hat{\boldsymbol{\Sigma}}_{11}^{-1/2}\mathbf{Y}_1\boldsymbol{\Delta}_1\mathbf{C}_T\\
\frac{\partial\|\mathbf{K}_{\mathcal{DCTW}}\|_*}{\partial\mathbf{Y}_1} &= \frac{1}{T-1}\left(\mathbf{F}^{(\text{pos})} - \mathbf{F}^{(\text{neg})}\right).
\end{aligned} \tag{13}$$

At this point, it is clear that CTW is a special case of DCTW. In fact, we arrive at CTW (Sec. 3.2) by simply considering a network with one layer. In this case, by setting $f_i(\mathbf{X}_i;\theta_i) = \mathbf{W}_i^\top\mathbf{X}_i$, Eq. 11 becomes equivalent to Eq. 10, while solving Eq. 12 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) [14, 26]. In more detail, instead of Eq. 12 we now optimise

$$\begin{aligned}
&\sum_{i,j=1}^{m} \operatorname{corr}(f_i(\mathbf{X}_i;\theta_i)\boldsymbol{\Delta}_i, f_j(\mathbf{X}_j;\theta_j)\boldsymbol{\Delta}_j)\\
&= \sum_{i,j}^{m} \left\|\mathbf{K}_{\mathcal{DCTW}}^{ij}\right\|_*.
\end{aligned} \tag{14}$$

where m is the number of sequences and $\mathbf{K}_{\mathcal{DCTW}}^{ij} = \hat{\boldsymbol{\Sigma}}_{ii}^{-1/2}\hat{\boldsymbol{\Sigma}}_{ij}\hat{\boldsymbol{\Sigma}}_{jj}^{-1/2}$. The subgradient of Eq. 14 can be computed in a straightforward manner by utilising Eq. 13. Note that by setting $\boldsymbol{\Delta}_i = \mathbf{I}$, Eq. 14 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 $\boldsymbol{\Delta}_i$ (e.g., [38]). DCTW is illustrated in Fig. 1.

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) [29] which finds the optimal alignment path given that the sequences reside in the same manifold (as explained in Sec. 3.2).
- Iterative Motion Warping (IMW) [17] 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) [37] as posed in section Sec. 3.2, CTW finds the optimal reduced dimensionality subspace such that the sequences are maximally linearly correlated.
- Generalized Time Warping (GTW) [38] 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 [34] 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 Sec. 5.2, where multiple feature sets are aligned , *(ii)* the MMI Facial Expression database Sec. 5.3, where we apply DCTW on the alignment of facial Action Units, *(iii)* the XRMB database Sec. 5.4 where we align acoustic and articulatory recordings, and finally *(iv)* the CUAVE database Sec. 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 [38]. 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 $\mathbf{p}_i \in \left\{x \in \mathbb{N}^{l_{\text{alg}}} | 1 \le x \le n_m\right\}$ is the alignment path for the ith sequence with a length l_{alg}. The error is then defined as

$$\begin{aligned}
\text{Err} &= \frac{\text{dist}(\mathbf{P}^{\text{alg}}, \mathbf{P}^{\text{ground}}) + \text{dist}(\mathbf{P}^{\text{ground}}, \mathbf{P}^{\text{alg}})}{l_{\text{alg}} + l_{\text{ground}}},\\
\text{dist}\left(\mathbf{P}^1, \mathbf{P}^2\right) &= \sum_{i=1}^{l_1} \min_{j=1}^{l_2} \left\|\mathbf{p}_{(i)}^1 - \mathbf{p}_{(j)}^2\right\|_2.
\end{aligned}$$

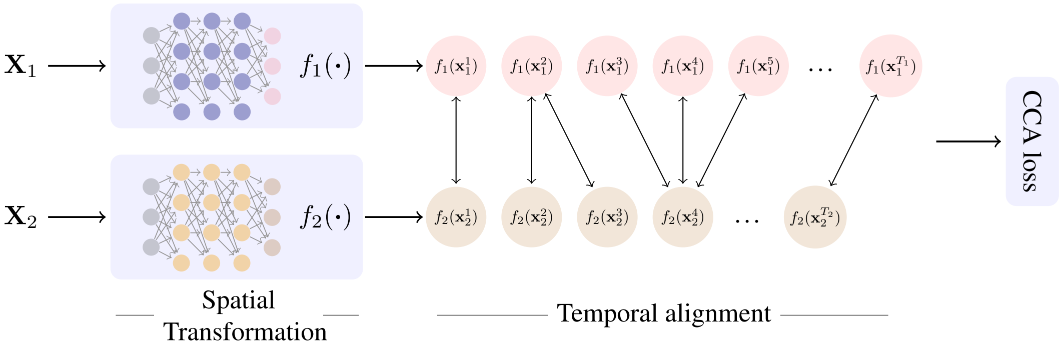


Figure 1: Illustration of the DCTW architecture with two networks, one for each temporal sequence. The model is trained end-to-end, first performing a spatial transformation of the data samples and then a temporal transformation such as the temporal sequences are maximally correlated.

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 [33]. We utilise full-batch optimisation with AdaGrad [9] 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) [22], 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 Sec. 5.2 where a smaller network was sufficient) we utilised a fixed three layer 200–100–100 fully connected topology, thus reducing the number the number of free hyperparameters of the architecture. : 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 [13], containing videos of nine subjects performing one of ten actions (e.g., walking). We adopt the experimental protocol described in [38], where 3 different shape features are computed for each sequence, namely *(1)* a binary mask, *(2)* Euclidean distance transform [24], and *(3)* the solution of the Poisson equation [12, 38]. 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 [38] are used. Following [37], [38], 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.

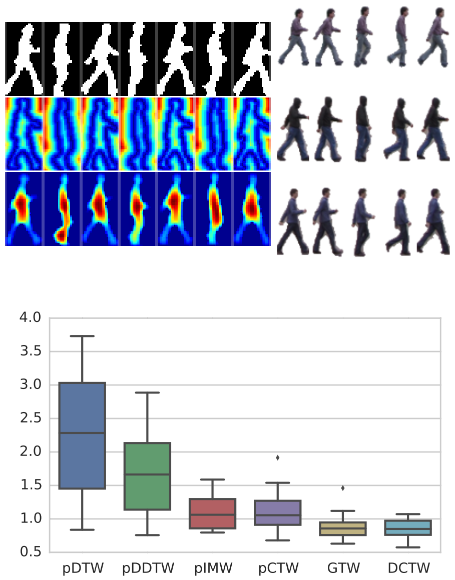


Figure 2: Aligning sequences of subjects performing similar actions from the Weizmann database. (left) the three computed features for each of the sequences (1) binary (2) euclidean (3) poisson solution. (middle) The aligned sequences using DCTW. (right) Alignment errors for each of the six techniques.

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 Fig. 2 we show the alignment error for ten randomly generated sets of videos. As DTW, DDTW, IMW, and CTW are only formulated for performing alignment between two sequences we use their multi-sequence extension as formulated in [39] 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 [27] 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 [19] 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 Fig. 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 Fig. 3 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 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.

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) [36]. 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) [7] 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 [36], 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.

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.

Results are presented in Tab. 1. Note that DCTW outperforms compared methods by a much larger margin than

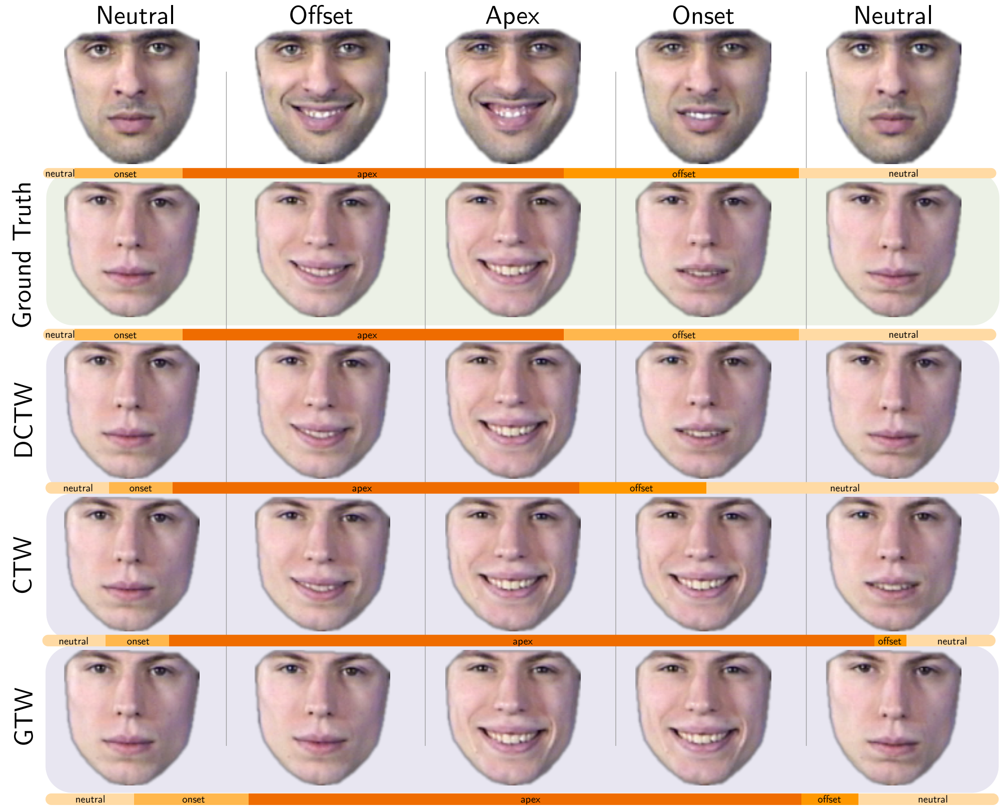


Figure 3: Facial expression alignment of videos S002–005 and S014–009 from MMI dataset (Sec. 5.3). Depicted frames for each temporal phase with duration $[t_s, t_e]$ correspond to the middle of each of the temporal phase, $t_c = \lceil \frac{t_s + t_e}{2} \rceil$. We also plot the temporal phases (● 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.

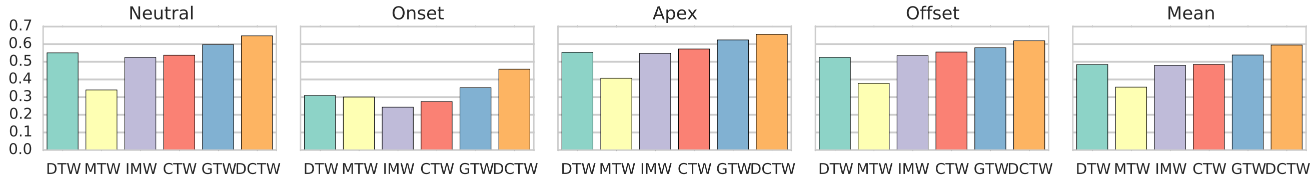


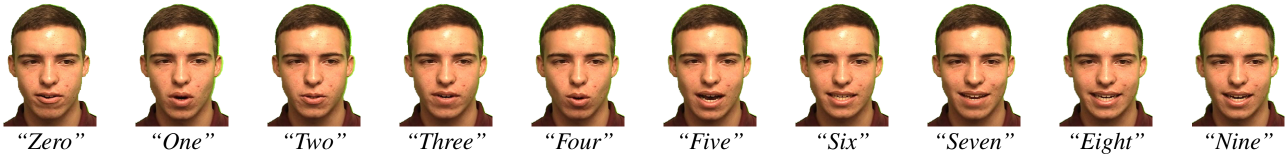
Figure 4: Temporal phase detection accuracy as defined by the ratio of correctly aligned frames with respect to the total duration for each temporal phase – the higher the better.

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.

5.5. Real Data IV: Alignment of Audio and Visual Streams

In our last (and arguably, most challenging) experiment, we aim to align the subject's visual and auditory utterances. To this end, we use the CUAVE [28] 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 Sec. 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 [19]. 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



“Zero” *“One”* *“Two”* *“Three”* *“Four”* *“Five”* *“Six”* *“Seven”* *“Eight”* *“Nine”*

Figure 5: The alignment results of DCTW for subject #1 of the CUAVE database.

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 outperforms the rest of the temporal alignment methods by a large margin. Again, the justification is similar to Sec. 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.

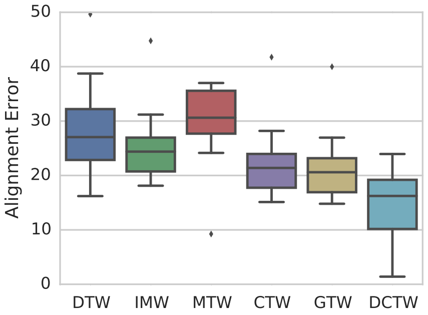


Figure 6: Alignment errors on the task of audio-visual temporal alignment. Note that videos better illustrating the results are contained in our supplementary material.

6. Computational details and discussion

Currently the computational cost of aligning a set of m sequences each of length of T_i samples each is $\mathcal{O}(\sum_{i,j}^{m} T_i T_j + eig(\sum_i d_i))$, which is bounded by the cost of performing DTW and secondly the cost of the singular value decomposition to calculate the derivatives in Eq. 13. As the decomposition is performed on the last layer of the network, which is of reduced dimensionality (100 units in our case) it is very cheap to compute in practise. In contrast other non-linear warping algorithms [34] require an expensive k-nearest neighbour and an extra eigendecomposition step or in the case of CTW [37] an eigendecomposition on the original correlation matrix which becomes much more expensive when dealing with data of high dimensionality.

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 [30] or GTW [38] 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. By means of various experiments on four real datasets, the significance of DCTW on multiple applications is highlighted, as the proposed method outperforms, in many cases by a very large margin, compared state-of-the-art methods for temporal alignment.

8. Acknowledgements

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) and by the 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). We would like to thank the NVIDIA Corporation for donating a Tesla K40 GPU used in this work.

References

[1] J. Aach and G. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17:495–508, 2001. 1

[2] G. Andrew et al. Deep Canonical Correlation Analysis. In *ICML*, volume 28, 2013. 2, 3, 4, 5, 6

[3] F. Bach. Consistency of trace norm minimization. *JMLR*, 9:1019–1048, 2008. 3, 4

[4] F. Bach and M. Jordan. A probabilistic interpretation of canonical correlation analysis. 2005. 2

[5] A. Bruderlin and L. Williams. Motion signal processing. In *SIGGRAPH*, pages 97–104, 1995. 1

[6] Y. Caspi and M. Irani. Aligning non-overlapping sequences. *IJCV*, 48:39–51, 2002. 1

[7] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE TASSP*, 28, 1980. 6

[8] F. De La Torre. A least-squares framework for component analysis. *IEEE TPAMI*, 34:1041–1055, 2012. 2

[9] J. Duchi et al. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *JMLR*, 12:2121–2159, 2011. 5

[10] R. Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE CVPR*, pages 580–587. IEEE, 2014. 2

[11] D. Gong and G. Medioni. Dynamic Manifold Warping for view invariant action recognition. In *IEEE CVPR*, pages 571–578, 2011. 1

[12] L. Gorelick et al. Shape representation and classification using the poisson equation. *IEEE TPAMI*, 28:1991–2004, 2006. 5

[13] L. Gorelick et al. Actions as space-time shapes. *IEEE TPAMI*, 29:2247–2253, 2007. 5

[14] M. Hasan. On multi-set canonical correlation analysis. In *IJCNN*, pages 1128–1133. IEEE, 2009. 1, 4

[15] 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. 2

[16] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006. 2

[17] E. Hsu, K. Pulli, and J. Popović. Style translation for human motion. In *SIGGRAPH*, volume 24, page 1082, 2005. 4

[18] 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. 1

[19] V. Kazemi and S. Josephine. One Millisecond Face Alignment with an Ensemble of Regression Trees. In *IEEE CVPR*, 2014. 6, 7

[20] 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. 2

[21] A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 2

[22] A. Maas et al. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, volume 30, 2013. 5

[23] K. Mardia et al. *Multivariate analysis*. Academic press, 1979. 2

[24] 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. 5

[25] J. Ngiam, A. Khosla, and M. Kim. Multimodal deep learning. *ICML*, 2011. 2

[26] A. A. Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE TIP*, 11(3):293–305, 2002. 4

[27] M. Pantic et al. Web-based database for facial expression analysis. In *ICME*, volume 2005, pages 317–321, 2005. 6

[28] E. Patterson et al. CUAVE: A new audio-visual database for multimodal human-computer interface research. *ICASSP*, 2:II–2017–II–2020, 2002. 7

[29] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*, volume 103. 1993. 1, 3, 4

[30] S. Salvador and P. Chan. Fastdtw: Toward accurate dynamic time warping in linear time and space. In *KDD-Workshop*, 2004. 8

[31] S. Shariat and V. Pavlovic. Isotonic CCA for sequence alignment and activity recognition. In *ICCV*, pages 2572–2578, 2011. 1

[32] Y. Taigman et al. Deepface: Closing the gap to human-level performance in face verification. In *IEEE CVPR*, pages 1701–1708. IEEE, 2014. 2

[33] P. Vincent et al. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *JMLR*, 11:3371–3408, 2010. 5

[34] H. Vu et al. Manifold Warping: Manifold Alignment over Time. In *AAAI*, 2012. 1, 4, 8

[35] A. Wang, J. Lu, G. Wang, J. Cai, and T.-J. Cham. Multimodal unsupervised feature learning for RGB-D scene labeling. In *ECCV*, pages 453–467. Springer, 2014. 2

[36] J. Westbury et al. X-ray microbeam speech production database. *JASA*, 88(S1):S56—-S56, 1990. 6

[37] F. Zhou and F. De La Torre. Canonical time warping for alignment of human behavior. *NIPS*, 2009. 1, 3, 4, 5, 8

[38] F. Zhou and F. De La Torre. Generalized time warping for multi-modal alignment of human motion. In *IEEE CVPR*, pages 1282–1289, 2012. 1, 4, 5, 8

[39] F. Zhou and F. De La Torre. Generalized Canonical Time Warping. *IEEE TPAMI*, 2015. 6