Dynamic Probabilistic CCA for Analysis of Affective Behaviour

Mihalis A. Nicolaou[1], Vladimir Pavlovic[2] and Maja Pantic[1,3]
{mihalis,m.pantic}@imperial.ac.uk http://ibug.doc.ic.ac.uk
vladimir@cs.rutgers.edu http://seqam.rutgers.edu

[1] Dept. of Computing, Imperial College London, UK
[2] Dept. of Computer Science, Rutgers University, USA
[3] EEMCS, University of Twente, Netherlands

Abstract. Fusing multiple continuous expert annotations is a crucial problem in machine learning and computer vision, particularly when dealing with uncertain and subjective tasks related to affective behaviour. Inspired by the concept of inferring shared and individual latent spaces in probabilistic CCA (PCCA), we firstly propose a novel, generative model which discovers temporal dependencies on the shared/individual spaces (DPCCA). In order to accommodate for temporal lags which are prominent amongst continuous annotations, we further introduce a latent warping process. We show that the resulting model (DPCTW) (i) can be used as a unifying framework for solving the problems of temporal alignment and fusion of multiple annotations in time, and (ii) that by incorporating dynamics, modelling annotation/sequence specific biases, noise estimation and time warping, DPCTW outperforms state-of-the-art methods for both the aggregation of multiple, yet imperfect expert annotations as well as the alignment of affective behavior.

1 Introduction

Most supervised learning tasks in computer vision and machine learning assume the existence of a reliable, objective label which corresponds to a given training instance. Nevertheless, especially in problems related to human behaviour, the annotation process (typically performed by multiple experts to reduce individual bias) can lead to inaccurate, ambiguous and subjective labels which in turn are used to train ill-generalisable models. Such problems arise not only due to the subjectivity of human annotators but also due to the fuzziness of the meaning associated with various labels related to human behaviour. The issue becomes even more prominent when the task is temporal, as it renders the labelling procedure vulnerable to temporal lags caused by varying response times of annotators. Considering that in many of the aforementioned problems the annotation is in a continuous real space (as opposed to discrete labels), the subjectivity of the annotators becomes much more difficult to model and fuse into a single "ground truth".

A recent emerging trend in affective computing is the adoption of real-valued, continuous dimensional emotion descriptions for learning tasks [1]. The space typically consists of two dimensions: valence (unpleasant to pleasant) and arousal (relaxed to aroused). In this description, each emotional state is mapped to a point in the valence/arousal space, thus overcoming the limitation of confining in a small set of discrete classes (such as the typically used six basic emotion classes). In this way, the expressiveness of the description is extended to non-basic emotions, typically manifested in everyday life (e.g., boredom). Nevertheless, the annotation of such data, although performed by multiple trained experts, results in labels which exhibit an amalgam of the aforementioned issues ([2], Fig. 1), leading researchers to adopt solutions based on simple averaging, reliance on a single annotator or quantising the continuous space and thus shifting the problem to the discrete domain (c.f. [3,4]).

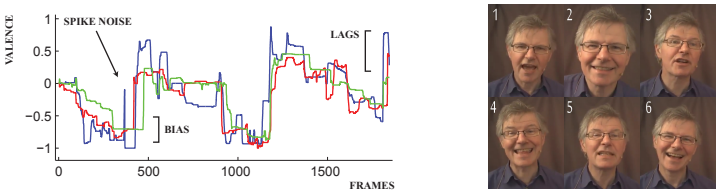

Fig. 1. Example valence annotations along with stills from the sequence.

A state-of-the-art approach in fusing multiple continuous annotations that *can* be applied to emotion descriptions is presented by Raykar et al. [5]. In this work, each noisy annotation is considered to be generated by a Gaussian distribution with the mean being the true label and the variance representing the annotation noise.

A main drawback of [5] lies in the assumption that the temporal correspondences of samples are known. One way to find such arbitrary temporal correspondences is with time warping. A state-of-the-art approach for time warping, Canonical Time Warping (CTW) [6], combines Dynamic Time Warping (DTW) and CCA with the aim of aligning a pair of sequences of both different duration and of different dimensionality. CTW accomplishes this by simultaneously finding the most correlated features and samples among the two sequences, both in feature space and time. This task is reminiscent of the goal of fusing annotations of two experts. However, CTW alignment does not directly yield the prototypical sequence, which is considered as a common, denoised and fused version of multiple experts' annotations. As a consequence, this renders neither of the two state-of-the-art methods applicable to our setting.

The latter observation precisely motivates our work; inspired by Probabilistic Canonical Correlation Analysis (PCCA) [7], we initially present the first generalisation of PCCA to learning temporal dependencies in the shared/individual spaces (DPCCA). By further augmenting DPCCA with time warping, the re-

sulting model (DPCTW) can be seen as a unifying framework, concisely applied to both problems. The individual contributions of this work can be summarised as follows:

- In comparison to state-of-the-art approaches in both fusion of multiple annotations and sequence alignment, our model has several advantages. We assume that the "true" annotation/sequence lies in a shared latent space. E.g., in the problem of fusing multiple emotion annotations, we know that the experts have a common training in annotation. Nevertheless, each carries a set of individual factors which can be assumed to be uninteresting (e.g., annotator/sequence specific bias). In the proposed model, individual factors are accounted for within an annotator-specific latent space, thus effectively preventing the contamination of the shared space by individual factors. Most importantly, we introduce latent-space dynamics which model temporal dependencies in both common and individual signals. Furthermore, due to the probabilistic and dynamic nature of the model, each annotator/sequence's uncertainty can be estimated for each *sample*, rather than for each sequence.
- In contrast to current work on fusing multiple annotations, we propose a novel framework able to handle temporal tasks. In addition to introducing dynamics, we also employ temporal alignment in order to eliminate temporal discrepancies amongst the annotations.
- Compared to state-of-the-art work on sequence alignment (e.g., CTW), we generalise traditional pairwise alignment approaches such as CTW to a multiple-sequence setting. We accomplish this by treating the problem in a generative probabilistic setting, both in the static (multiset PCCA) and dynamic case (Dynamic PCCA).

The rest of the paper is organised as follows: In Section 2, we describe PCCA and present our extension to multiple sequences. In Section 3, we introduce our Dynamic PCCA, which we subsequently extend with latent space time-warping as described in Section 4. In Section 5, we present various experiments on both synthetic (Sec. 5.1) and real (Sec. 5.2, 5.3) experimental data, emphasising the advantages of the proposed methods on both the fusion of multiple annotations and sequence alignment.

2 Multiset Probabilistic CCA

We consider the probabilistic interpretation of CCA, introduced by Bach & Jordan [8] and generalised by Klami & Kaski [7]. In this section, we present an extended version of PCCA (multiset PCCA[1]) [7] which is able to handle any arbitrary number of sets. We consider a collection of datasets $\mathcal{D} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N\}$, with each $\mathbf{X}_i \in \mathbb{R}^{D_i \times T}$. By adopting the generative model for PCCA, the observation sample n of set $\mathbf{X}_i \in \mathcal{D}$ is assumed to be generated as:

$$\mathbf{x}_{i,n} = f(\mathbf{z}_n|\mathbf{W}_i) + g(\mathbf{z}_{i,n}|\mathbf{B}_i) + \epsilon_i, \tag{1}$$

[1] For simplicity, in the following sections we refer to multiset PCCA as PCCA.

where $\mathbf{Z}_i = [\mathbf{z}_{i,1}, \ldots, \mathbf{z}_{i,t}]$ and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_t]$ are the *independent* latent variables that capture the set-specific individual characteristics and the shared signal amongst all observation sets respectively. $f(.)$ and $g(.)$ are functions that transform each of the latent signals $\mathbf{Z}$ and $\mathbf{Z}_i$ into the observation space. They are parametrised by $\mathbf{W}_i$ and $\mathbf{B}_i$, while the noise for each set is represented by ϵ_i, with $\epsilon_i \perp \epsilon_j, i \neq j$. Similarly to [7], $\mathbf{z}_n$, $\mathbf{z}_{i,n}$ and ϵ_i are considered to be independent (both over the set and the sequence) and normally distributed:

$$\mathbf{z}_n, \mathbf{z}_{i,n} \sim \mathcal{N}(0, \mathbf{I}), \epsilon_i \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I}). \tag{2}$$

By considering f and g to be linear functions we have $f = \mathbf{W}_i \mathbf{z}_n$ and $g = \mathbf{B}_i \mathbf{z}_{i,n}$, transforming the model presented in Eq. 1, to:

$$\mathbf{x}_{i,n} = \mathbf{W}_i \mathbf{z}_n + \mathbf{B}_i \mathbf{z}_{i,n} + \epsilon_i. \tag{3}$$

Learning the multiset PCCA can be accomplished by generalising the EM algorithm presented in [7], applied to two or more sets. Firstly, $P(\mathcal{D}|\mathbf{Z}, \mathbf{Z}_1, \ldots, \mathbf{Z}_N)$ is marginalised over set-specific factors $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ and optimised on each $\mathbf{W}_i$. This leads to the generative model $P(\mathbf{x}_{i,n}|\mathbf{z}_n) \sim \mathcal{N}(\mathbf{W}_i \mathbf{z}_n, \mathbf{\Psi}_i)$, where $\mathbf{\Psi}_i = \mathbf{B}_i \mathbf{B}_i^T + \sigma_i^2 \mathbf{I}$. Subsequently, $P(\mathcal{D}|\mathbf{Z}, \mathbf{Z}_1, \ldots, \mathbf{Z}_N)$ is marginalised over the common factor $\mathbf{Z}$ and then optimised on each $\mathbf{B}_i$ and σ_i. When generalising the algorithm for more than two sets, we also have to consider how to (i) obtain the expectation of the latent space and (ii) provide stable variance updates for all sets.

Two quantities are of interest regarding the latent space estimation. The first is the common latent space given a set, $\mathbf{Z}|\mathbf{X}_i$. In the classical CCA this is analogous to finding the canonical variables [7]. We estimate the posterior of the shared latent variable $\mathbf{Z}$ as follows:

$$P(\mathbf{z}_n|\mathbf{x}_{i,n}) \sim \mathcal{N}(\boldsymbol{\gamma}_i \mathbf{x}_{i,n}, \mathbf{I} - \boldsymbol{\gamma}_i \mathbf{W}_i), \boldsymbol{\gamma}_i = \mathbf{W}_i^T (\mathbf{W}_i \mathbf{W}_i^T + \mathbf{\Psi}_i)^{-1}. \tag{4}$$

The latent space given the n-th sample from *all* sets in $\mathcal{D}$, which provides a better estimate of the shared signal manifested in all observation sets is estimated as

$$P(\mathbf{z}_n|\mathbf{x}_{1:N,n}) \sim \mathcal{N}(\boldsymbol{\gamma}\mathbf{X}, \mathbf{I} - \boldsymbol{\gamma}\mathbf{W}), \boldsymbol{\gamma} = \mathbf{W}^T (\mathbf{W}\mathbf{W}^T + \mathbf{\Psi})^{-1}, \tag{5}$$

while the matrices $\mathbf{W}$, $\mathbf{\Psi}$ and $\mathbf{X}_n$ are defined as $\mathbf{W} = [\mathbf{W}_1; \ldots; \mathbf{W}_N]$, $\mathbf{\Psi}$ as the block diagonal matrix of $\mathbf{\Psi}_i$ and $\mathbf{X}_n = [\mathbf{X}_{i,n}^T; \ldots; \mathbf{X}_{N,n}^T]$. Finally, the variance is recovered on the full model, $x_{i,n} \sim \mathcal{N}(\mathbf{W}_i \mathbf{z}_n + \mathbf{B}_i \mathbf{z}_{i,n}, \sigma_i^2 \mathbf{I})$, as

$$\sigma_i^2 = trace(\mathbf{S} - \mathbf{X}\mathbb{E}[\mathbf{Z}^T|\mathbf{X}]\mathbf{C}^T - \mathbf{C}\mathbb{E}[\mathbf{Z}|\mathbf{X}]\mathbf{X}^T - \mathbf{C}\mathbb{E}[\mathbf{Z}\mathbf{Z}^T|\mathbf{X}]\mathbf{C}^T)_i \frac{T}{D_i}, \tag{6}$$

where $\mathbf{S}$ is the sample covariance matrix, D_i is the dimensionality of the samples in set $\mathbf{X}_i$, $\mathbf{B}$ is the block diagonal matrix of $\mathbf{B}_{i=1:N}$ and $\mathbf{C} = [\mathbf{W}|\mathbf{B}]$. We denote that the subscript i refers to the i-th block of the full covariance matrix.

3 Dynamic PCCA (DPCCA)

The PCCA model described in Section 2 exhibits several advantages when compared to the classical formulation of CCA, mainly by providing a probabilistic

estimation of a latent space shared by an arbitrary collection of datasets along with explicit noise estimation. Nevertheless, static models are unable to learn temporal dependencies which are very likely to exist when dealing with real-life problems. In fact, dynamics are deemed essential for successfully performing tasks such as emotion recognition, AU detection etc. [9].

Motivated by the former observation, we propose a dynamic generalisation of the static PCCA model introduced in the previous section, where we now treat each $\mathbf{X}_i$ as a temporal sequence. For simplicity of presentation, we introduce a linear model[2] where Markovian dependencies are learnt in the latent spaces $\mathbf{Z}$ and $\mathbf{Z}_i$. In other words, the variable $\mathbf{Z}$ models the temporal, shared signal amongst all observation sequences, while $\mathbf{Z}_i$ captures the temporal, individual characteristics of each sequence. It is easy to observe that such a model fits perfectly with the problem of fusing of multiple annotators, as it does not only capture the temporal shared signal of all annotators, but also models the unwanted, annotator-specific factors over time.

Essentially, instead of directly applying the doubly independent priors to $\mathbf{Z}$ as in Eq. 2, we now use the following:

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}) \sim \mathcal{N}(\mathbf{A}_z\mathbf{z}_{t-1}, \mathbf{V}_Z)), \tag{7}$$

$$p(\mathbf{z}_{i,t}|\mathbf{z}_{i,t-1}) \sim \mathcal{N}(\mathbf{A}_{z_i}\mathbf{z}_{i,t-1}, \mathbf{V}_{Z_i}), n = 1, \ldots, N, \tag{8}$$

where the transition matrices $\mathbf{A}_z$ and $\mathbf{A}_{z_i}$ model the latent space dynamics for the shared and sequence-specific space respectively. Thus, idiosyncratic characteristics of dynamic nature appearing in a single sequence can be accurately estimated and prevented from contaminating the estimation of the shared signal.

The resulting model bears similarities with traditional Linear Dynamic System (LDS) models (e.g. [12]) and the so-called Factorial Dynamic Models, c.f. [13]. Along with Eq. 7,8 and noting Eq. 3, the dynamic, generative model for DPCCA[3] can be described as

$$\mathbf{x}_{i,t} = \mathbf{W}_{i,t}\mathbf{z}_t + \mathbf{B}_i\mathbf{z}_{i,t} + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma_i^2\mathbf{I}), \tag{9}$$

where $\mathbf{x}_{i,t}, \mathbf{z}_{i,t}$ refer to the i-th observation sequence, timestep t.

3.1 Inference

To perform inference, we reduce the DPCCA model to a LDS. This can be accomplished by defining a joint space $\hat{\mathbf{Z}} = [\mathbf{Z}; \mathbf{Z}_1; \ldots; \mathbf{Z}_N]$ with parameters $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{W}, \mathbf{B}, \mathbf{V}_{\hat{z}}, \hat{\boldsymbol{\Sigma}}\}$. Dynamics in this joint space are described as $\mathbf{X}_t =$

[2] A non-linear DPCCA model can be derived similarly to [10,11].

[3] As can be easily seen, the model by Raykar et al. [5] can be considered as a special case of DPCCA, by setting $\mathbf{W} = \mathbf{I}$, $\mathbf{B} = \mathbf{0}$ and disregarding dynamics.

$[\mathbf{W}|\mathbf{B}]\hat{\mathbf{Z}}_t + \boldsymbol{\epsilon}, \hat{\mathbf{Z}}_t = \mathbf{A}\hat{\mathbf{Z}}_{t-1} + \mathbf{u}$, where the noise processes $\boldsymbol{\epsilon}$ and $\mathbf{u}$ are defined as

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \underbrace{\begin{bmatrix}\sigma_1^2\mathbf{I} & & \\ & \ddots & \\ & & \sigma_N^2\mathbf{I}\end{bmatrix}}_{\hat{\boldsymbol{\Sigma}}}\right), \mathbf{u} \sim \mathcal{N}\left(0, \underbrace{\begin{bmatrix}\mathbf{V}_z & & & \\ & \mathbf{V}_{z_1} & & \\ & & \ddots & \\ & & & \mathbf{V}_{z_N}\end{bmatrix}}_{\mathbf{V}_{\hat{z}}}\right). \tag{10}$$

The matrices used above are defined as $\mathbf{X} = [\mathbf{X}_1; \ldots; \mathbf{X}_N]$, $\mathbf{W} = [\mathbf{W}_1; \ldots; \mathbf{W}_N]$, $\mathbf{B}$ as the block diagonal matrix of $[\mathbf{B}_1, \ldots, \mathbf{B}_N]$ and finally, $\mathbf{A}$ as the block diagonal matrix of $[\mathbf{A}_z, \mathbf{A}_{z_1}, \ldots, \mathbf{A}_{z_N}]$. Similarly to LDS, the joint log-likelihood function of DPCCA is defined as

$$lnP(\mathbf{X}, \mathbf{Z}|\theta) = lnP(\hat{\mathbf{z}}_1|\mu, V) + \sum_{t=2}^{T} lnP(\hat{\mathbf{z}}_t|\mathbf{A}, \mathbf{V}_{\hat{z}}) + \sum_{t=1}^{T} lnP(\mathbf{x}_t|\hat{\mathbf{z}}_t, \mathbf{W}, \mathbf{B}, \hat{\boldsymbol{\Sigma}}). \tag{11}$$

In order estimate the latent spaces, we apply the Rauch-Tung-Striebel (RTS) smoother. In this way, we obtain $\mathbb{E}[\hat{\mathbf{z}}_t|\mathbf{X}^T]$, $V[\hat{\mathbf{z}}_t|\mathbf{X}^T]$ and $V[\hat{\mathbf{z}}_t\hat{\mathbf{z}}_{t-1}|\mathbf{X}^T]$[4].

3.2 Parameter Estimation

The parameter estimation of the M-step has to be derived specifically for this factorised model. We consider the expectation of the joint model log-likelihood (Eq. 11) w.r.t. posterior and obtain the partial derivatives of each parameter for finding the stationary points. Note the $\mathbf{W}$ and $\mathbf{B}$ matrices appear in the likelihood as:

$$\mathbb{E}_{\hat{z}}[lnP(\mathbf{X}, \hat{\mathbf{Z}})] = -\frac{T}{2} ln|\hat{\boldsymbol{\Sigma}}| - \mathbb{E}_{\hat{z}}\left[\sum_{t=1}^{T} \left(\mathbf{x}_t - [\mathbf{W}|\mathbf{B}]\hat{\mathbf{z}}_t\right)^T \hat{\boldsymbol{\Sigma}}^{-1} \left(\mathbf{x}_t - [\mathbf{W}|\mathbf{B}]\hat{\mathbf{z}}_t\right)\right] + \ldots. \tag{12}$$

Since they are composed of individual $\mathbf{W}_i$ and $\mathbf{B}_i$ matrices (which are parameters for each sequence i), we calculate the partial derivatives $\partial\mathbf{W}_i$ and $\partial\mathbf{B}_i$ in Eq. 12. Subsequently, by setting to zero and re-arranging, we obtain the update equations for each $\mathbf{W}_i^*$ and $\mathbf{B}_i^*$:

$$\mathbf{W}_i^* = \left(\sum_{t=1}^{T} \mathbf{x}_{i,t}\mathbb{E}[\mathbf{z}_{i,t}] - \mathbf{B}_i^*\mathbb{E}[\mathbf{z}_{i,t}\mathbf{z}_t^T]\right)\left(\sum_{t=1}^{T} \mathbb{E}[\mathbf{z}_t\mathbf{z}_t^T]\right)^{-1} \tag{13}$$

$$\mathbf{B}_i^* = \left(\sum_{t=1}^{T} \mathbf{x}_{i,t}\mathbb{E}[\mathbf{z}_t^T] - \mathbf{W}_i^*\mathbb{E}[\mathbf{z}_t\mathbf{z}_{i,t}^T]\right)\left(\sum_{t=1}^{T} \mathbb{E}[\mathbf{z}_{i,t}\mathbf{z}_{i,t}^T]\right)^{-1} \tag{14}$$

[4] We denote that the complexity of RTS is cubic in the dimension of the state space. Thus, when estimating large-dimensionality latent spaces, computational or numerical (due to the inversion of large matrices) issues may arise. If any of the above is a concern, the complexity of RTS can be reduced to quadratic [14], while the inference can be performed more efficiently similarly to [13].

Note that the weights are *coupled* and thus the optimal solution should be found iteratively. As can be seen, in contrast to PCCA, in DPCCA the individual factors of each sequence are explicitly calculated instead of being marginalised out. In a similar fashion, the transition weight updates for the individual factors $\mathbf{Z}_i$ are as follows:

$$\mathbf{A}^*_{z,i} = \left(\sum_{t=2}^{T} E[\mathbf{z}_{i,t}\mathbf{z}_{i,t-1}^T]\right)\left(\sum_{t=2}^{T} E[\mathbf{z}_{i,t-1}\mathbf{z}_{i,t-1}^T]\right)^{-1} \tag{15}$$

where by removing the subscript i we obtain the updates for $\mathbf{A}_z$, corresponding to the shared latent space $\mathbf{Z}$. Finally, the noise updates $\mathbf{V}_{\hat{Z}}$ and $\hat{\mathbf{\Sigma}}$ are estimated similarly to LDS [12].

4 Dynamic Probabilistic CCA with Time Warping

Both PCCA and DPCCA exhibit several advantages in comparison to the classical formulation of CCA. Mainly, as we have shown, (D)PCCA can inherently handle more than two sequences, building upon the multiset nature of PCCA. This is in contrast to the classical formulation of CCA, which due to the pairwise nature of the correlation operator is limited to two sequences[5]. This is crucial for the problems at hand since both methods yield an accurate estimation of the underlying signals of *all* observation sequences, free of individual factors and noise. However, both PCCA and DPCCA carry the assumption that the temporal correspondences between samples of different sequences are *known*, i.e. that the annotation of expert i at time t directly corresponds to the annotation of expert j at the same time. Nevertheless, this assumption is often violated since different experts exhibit different time lags in annotating the same process (e.g., Fig. 1, [15]). Motivated by the latter, we extend the DPCCA model to account for this *misalignment* of data samples by introducing a latent warping process into DPCCA, in a manner similar to [6]. In what follows, we firstly describe some basic background on time-warping and subsequently proceed to define our model.

4.1 Time Warping

Dynamic Time Warping (DTW)[16] is an algorithm for optimally aligning two sequences of possibly different lengths. Given sequences $\mathbf{X} \in \mathbb{R}^{d\times T_x}$ and $\mathbf{Y} \in \mathbb{R}^{d\times T_y}$, DTW aligns the samples of each sequence by minimising the sum-of-squares cost, i.e. $||\mathbf{X}\mathbf{\Delta}_x^T - \mathbf{Y}\mathbf{\Delta}_y^T||_F^2$, where $\mathbf{\Delta}_x$ and $\mathbf{\Delta}_y$ are binary selection matrices, effectively re-mapping the samples of each sequence. Although the number of possible alignments is exponential in T_xT_y, employing dynamic programming can recover the optimal path in $\mathcal{O}(T_xT_y)$. Furthermore, the solution must satisfy the boundary, continuity and monotonicity constraints.

An important limitation of DTW is the inability to align signals of different dimensionality. Motivated by the former, CTW [6] combines CCA and DTW,

[5] Extended CCA can handle multiple sequences but this involves pairwise averaging.

thus alowing the alignment of signals of different dimensionality by projecting into a common space via CCA. The optimisation function now becomes $||\mathbf{V}_x^T\mathbf{X}\mathbf{\Delta}_x^T - \mathbf{V}_y^T\mathbf{Y}\mathbf{\Delta}_y^T||_F^2$, where $\mathbf{X} \in \mathbb{R}^{d_x \times T_x}, \mathbf{Y} \in \mathbb{R}^{d_y \times T_x}$, and $\mathbf{V}_x, \mathbf{V}_y$ are the projection operators (matrices).

4.2 DPCTW Model

We define DPCTW based on the graphical model presented in Fig. 2. Given a set $\mathcal{D}$ of N sequences, with each sequence $\mathbf{X}_i = [\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,T_i}]$, we postulate the latent common Markov process $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_t\}$. Firstly, $\mathbf{Z}$ is warped using the warping operator $\mathbf{\Delta}_i$, resulting in the warped latent sequence $\boldsymbol{\zeta}_i$. Subsequently, each $\boldsymbol{\zeta}_i$ generates each observation sequence $\mathbf{X}_i$, also considering the annotator/sequence bias $\mathbf{Z}_i$ and the observation noise σ_i^2. We note that we do not impose parametric models for warping processes. Inference in this general model can be prohibitively expensive, in particular because of the need to handle the unknown alignments. We instead propose to handle the inference in two steps: (i) fix the alignments $\mathbf{\Delta}_i$ and find the latent $\mathbf{Z}$ and $\mathbf{Z}_i$'s, and (ii) given the estimated $\mathbf{Z}, \mathbf{Z}_i$ find the optimal warpings $\mathbf{\Delta}_i$. For this, we propose to optimise the following objective function:

$$\mathcal{L}_{(D)PCTW} = \sum_{i}^{N} \sum_{j,j\neq i}^{N} \frac{1}{N(N-1)} ||\mathbb{E}[\mathbf{Z}|\mathbf{X}_i]\mathbf{\Delta}_i - \mathbb{E}[\mathbf{Z}|\mathbf{X_j}]\mathbf{\Delta_j}||_F^2 \tag{16}$$

where when using PCCA, $\mathbb{E}[\mathbf{Z}|\mathbf{X}_i] = \mathbf{W}_i^T(\mathbf{W}_i\mathbf{W}_i^T + \mathbf{\Psi}_i)^{-1}\mathbf{X}_i$ (Eq. 4). For DPCCA, $\mathbb{E}[\mathbf{Z}|\mathbf{X}_i]$ is inferred via RTS smoothing (Sec. 3). A summary of the full algorithm is presented in Algorithm 1.

Guide Tree Progressive Alignment. We note that the optimal solution for time warping can be found for any number of sequences. Nevertheless, the complexity of the problem becomes exponential with the increase of the number of sequences. Therefore, for more than 2 sequences, we adopt an approximation based on a variation of Progressive Alignment using a guide tree, adjusted to fit a continuous space. Similar algorithms are used in state-of-the-art sequence alignment software in biology, e.g., Clustar.

5 Experiments

To evaluate the proposed models, in this section, we present a set of experiments on both synthetic (Sec. 5.1) and real (Sec. 5.2 & 5.3) data.

5.1 Synthetic Data

For synthetic experiments, we employ a similar setting to [6]. A set of 2D spirals are generated as $\mathbf{X}_i = \mathbf{U}_i^T\tilde{\mathbf{Z}}\mathbf{M}_i^T + \mathbf{N}$, where $\tilde{\mathbf{Z}} \in \mathbb{R}^{2\times T}$ is the true latent signal

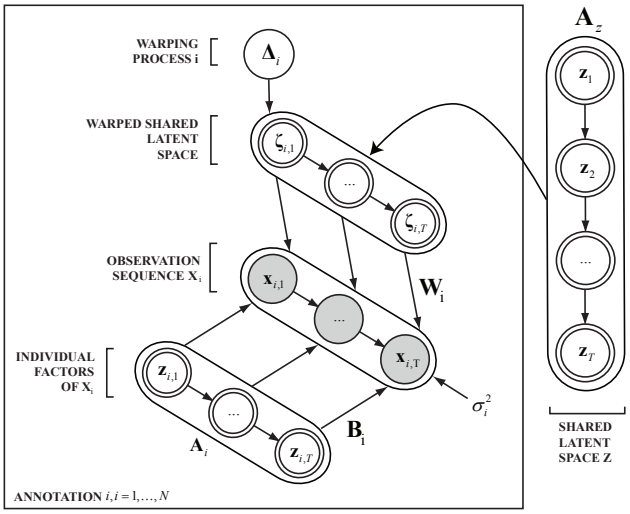

Fig. 2. Graphical model of DPCTW. Shaded nodes represent the observations. By ignoring the temporal dependencies, we obtain the PCTW model.

Algorithm 1: Dynamic Probabilistic CCA with Time Warpings

Data: $\mathbf{X}_1, \ldots, \mathbf{X}_N$
Result: $P(\mathbf{Z}|\mathbf{X}_1, \ldots \mathbf{X}_N), P(\mathbf{Z}|\mathbf{X}_i), \mathbf{\Delta}_i, \sigma_i^2$ where $i = 1, \ldots, N$
begin
  repeat
    $(\mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_N) \leftarrow$ time warping$(\mathbb{E}[\hat{\mathbf{z}}_t|\mathbf{X}_1^T], \ldots, \mathbb{E}[\hat{\mathbf{z}}_t|\mathbf{X}_N^T])^*$
    repeat
      Estimate $\mathbb{E}[\hat{\mathbf{z}}_t|\mathbf{X}^T]$, $V[\hat{\mathbf{z}}_t|\mathbf{X}^T]$ and $V[\hat{\mathbf{z}}_t\hat{\mathbf{z}}_{t-1}|\mathbf{X}^T]$ (RTS)
      for $i = 1, \ldots, N$ **do**
        repeat
          Update $\mathbf{W}_i^*$ according to Eq. 13
          Update $\mathbf{B}_i^*$ according to Eq. 14
        until $\mathbf{W}_i$, $\mathbf{B}_i$ *converge*
        Update $\mathbf{A}_i^*$ according to Eq. 15
      Update $\mathbf{A}^*, \mathbf{V}_{\hat{Z}}^*, \hat{\mathbf{\Sigma}}^*$ according to Sec. 3.2
    until *DPCCA converges*
    for $i = 1, \ldots, N$ **do**
      $\theta_i = \left\{ \begin{bmatrix} \mathbf{A}_z & 0 \\ 0 & \mathbf{A}_i \end{bmatrix}, \mathbf{W}_i, \mathbf{B}_i, \begin{bmatrix} \mathbf{V_Z} & 0 \\ 0 & \mathbf{V}_i \end{bmatrix}, \sigma_i^2 \mathbf{I} \right\}$
      Estimate $\mathbb{E}[\hat{\mathbf{z}}_t|\mathbf{X}_i^T]$, $V[\hat{\mathbf{z}}_t|\mathbf{X}_i^T]$ and $V[\hat{\mathbf{z}}_t\hat{\mathbf{z}}_{t-1}|\mathbf{X}_i^T]$ (RTS(θ_i))
  until $\mathcal{L}_{DPCTW}$ *converges*

* In the first iteration since $\mathbb{E}[\hat{\mathbf{z}}_t|\mathbf{X}_i^T]$ is not known, use $\mathbf{X}_i$ instead

which generates the $\mathbf{X}_i$, while the $\mathbf{U}_i \in \mathbb{R}^{2\times 2}$ and $\mathbf{M}_i \in \mathbb{R}^{T_i \times m}$ matrices impose random spatial and temporal warping. The signal is furthermore perturbed by additive noise via the matrix $\mathbf{N} \in \mathbb{R}^{2\times T}$. Each $\mathbf{N}(i,j) = e \times b$, where $e \sim \mathcal{N}(0,1)$ and b follows a Bernoulli distribution with $P(b = 1) = 1$ for Gaussian and $P(b = 1) = 0.4$ for spike noise.

This experiment can be interpreted as both of the problems we are examining. Viewed as a sequence alignment problem the goal is to recover the alignment of each noisy $\mathbf{X}_i$, where in this case the true alignment is known. Considering the problem of fusing multiple annotations, the latent signal $\tilde{\mathbf{Z}}$ represents the true annotation while the individual $\mathbf{X}_i$ form the set of noisy annotations containing annotation-specific characteristics. The goal is to recover the true latent signal (in DPCCA terms, $\mathbb{E}[\mathbf{Z}|\mathbf{X}_1, \ldots, \mathbf{X}_N]$).

For qualitative evaluation, in Fig. 3, we present an example of applying (D)PCTW on 5 sequences. As can be seen, DPCTW is able to recover the true, de-noised, latent signal which generated the noisy observations (Fig. 3(e)), while also aligning the noisy sequences (Fig. 3(c)). Due to the temporal modelling of DPCTW, the recovered latent space is almost identical to the true signal $\tilde{\mathbf{Z}}$ (Fig. 3(b)). PCTW on the other hand is unable to entirely remove the noise (Fig. 3(d)). Fig. 4 shows further results comparing related methods. For two sequences, CTW outperforms DTW as expected. PCTW is better than CTW (but marginally, in the case of spike noise). DPCTW provides much better alignment than other methods.

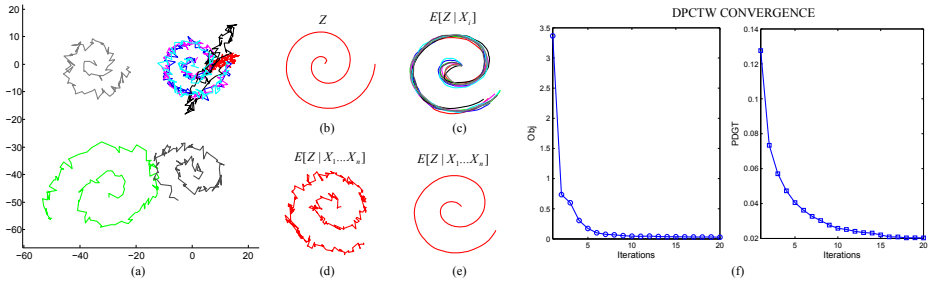

Fig. 3. Noisy synthetic data experiment. (a) Initial, noisy time series. (b) True latent signal from which the noisy, transformed spirals where attained in (a). (c) The alignment achieved by DPCTW. The latent space $E[\mathbf{Z}|\mathbf{X}_1, \ldots, \mathbf{X}_N]$ (i.e., the recovered shared latent signal) for (d) PCTW and (e) DPCTW. (f) Convergence of DPCTW in terms of the objective (Obj) and error wrt. ground truth (PDGT).

5.2 Real Data I: Fusing Multiple Annotations

In order to evaluate (D)PCTW in case of real data, we employ the SEMAINE database [15]. The database contains a set of audio-visual recordings of subjects interacting with operators. Each operator assumes a certain personality - happy, gloomy, angry and pragmatic - with a goal of inducing spontaneous emotions by

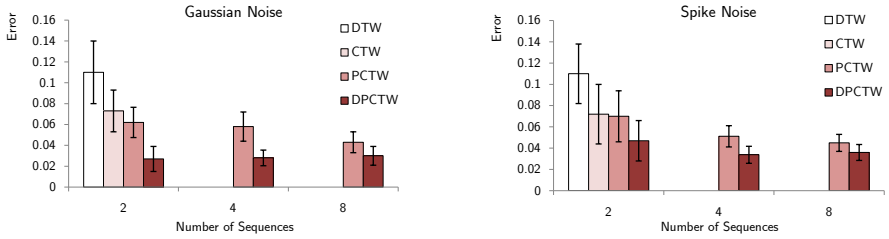

Fig. 4. Synthetic experiment comparing the alignment attained by DTW, CTW, PCTW and DPCTW on spirals with spiked and Gaussian noise.

the subject during a naturalistic conversation. We use a portion of the database containing recordings of 6 different subjects. As the database was annotated in terms of valence/arousal by a set of experts, no single ground truth is provided along with the recordings. Thus, by considering $\mathbf{X}$ to be the set of annotations in valence or arousal and applying (D)PCTW, we obtain $\mathbb{E}[\mathbf{Z}] \in \mathbb{R}^{1\times T}$ given all annotations, which represents the shared latent space with annotator-specific factors and noise removed. We assume that this expectation represents the ground truth. An example of this procedure for (D)PCTW can be found in Fig. 5. As can be seen, DPCTW provides a smooth, aligned estimate, eliminating temporal discrepancies, spike-noise and annotator bias.

To obtain features for evaluating the ground truth, we track the facial expressions of each subject by employing the Patras - Pantic particle filtering tracking scheme [17]. The tracked points include the corners of the eyebrows (4 points), the eyes (8 points), the nose (3 points), the mouth (4 points) and the chin (1 point), resulting in 20 2D points for each frame. For evaluation, we consider a training sequence $\mathbf{X}$, for which the set of annotations $\mathcal{A}_x = \{\mathbf{a}_1, \ldots, \mathbf{a}_R\}$ is known. From this set $\mathcal{A}_x$, we derive the ground truth $\mathcal{GT}_{\mathbf{X}}$ - for (D)PCTW, $\mathcal{GT}_{\mathbf{X}} = \mathbb{E}[\mathbf{Z}|\mathcal{A}_x]$. Using the tracked points $\mathcal{P}_{\mathbf{X}}$ for the sequence, we train a regressor to learn the function $f_x : \mathcal{P}_{\mathbf{X}} \rightarrow \mathcal{GT}_{\mathbf{X}}$. In (D)PCTW, $\mathcal{P}_x$ is firstly aligned with $\mathcal{GT}_x$ as they are not necessarily of equal length. Subsequently given a testing sequence $\mathbf{Y}$ with tracked points $\mathcal{P}_y$, using f_x we predict the valence/arousal ($f_x(\mathcal{P}_y)$). The procedure for deriving the ground truth is then applied on the annotations of sequence $\mathbf{Y}$, and the resulting $\mathcal{GT}_y$ is evaluated against $f_x(\mathcal{P}_y)$. The correlation of the aligned $\mathcal{GT}_y$ and $f_x(\mathcal{P}_y)$ is then used as the evaluation metric for all compared methods.

The reasoning behind this experiment is that the "best" ground truth should maximally correlate with the corresponding input features - thus enabling any regressor to learn the mapping function more accurately. For regression, we employ RVM [18] with a Gaussian kernel. We perform both session-dependent experiments, where the validation was performed on each session separately, and session-independent experiments where different sessions where used for training/testing. In this way, we validate the derived ground truth generalisation ability (i) when the set of annotators is the same and (ii) when the set of annotators may differ. The obtained results are presented in Table 1. As can be seen,

taking the average gives the worse results (as expected). The model of Raykar et al. [5] provides better results, as it estimates the variance of each annotator. Modelling annotator bias and noise with (D)PCCA further improves the results. It is important to note that incorporating alignment appears to be significant for deriving the ground truth; this is reasonable since when the annotations are misaligned, shared information may be modelled as individual factors or vice-versa. Finally, DPCTW provides the best results, confirming our assumption that combining dynamics, temporal alignment, modelling noise and individual-annotator bias leads to a more objective ground truth.

Table 1. Comparison of ground truth evaluation based on the correlation coefficient (COR), on session dependent (SD) and session independent (SI) experiments.

		DPCTW		**PCTW**		**DPCCA**		**PCCA**		**Raykar** [5]		**AVG**	
		COR	$\boldsymbol{\sigma}$	**COR**	$\boldsymbol{\sigma}$	**COR**	$\boldsymbol{\sigma}$	**COR**	$\boldsymbol{\sigma}$	**COR**	$\boldsymbol{\sigma}$	**COR**	$\boldsymbol{\sigma}$
SD	**Valence**	**0.77**	*0.18*	0.70	*0.19*	0.64	*0.21*	0.63	*0.20*	0.61	*0.20*	0.54	*0.36*
	Arousal	**0.75**	*0.22*	0.64	*0.22*	0.63	*0.23*	0.63	*0.26*	0.60	*0.25*	0.42	*0.41*
SI	**Valence**	**0.72**	*0.22*	0.66	*0.24*	0.62	*0.25*	0.58	*0.23*	0.57	*0.27*	0.53	*0.33*
	Arousal	**0.71**	*0.20*	0.61	*0.23*	0.59	*0.23*	0.52	*0.28*	0.50	*0.29*	0.33	*0.40*

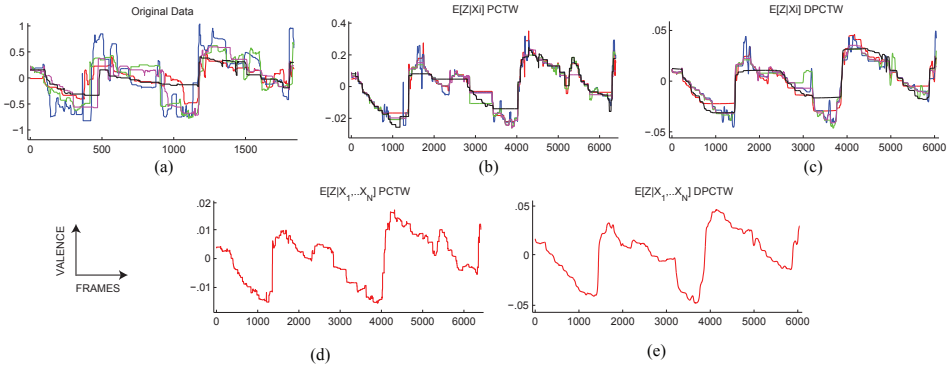

Fig. 5. Applying (D)PCTW to continuous emotion annotations. (a) Original valence annotations from 5 experts. (b,c) Alignment obtained by PCTW and DPCTW respectively, (d,e) Ground truth obtained by PCTW and DPCTW respectively.

5.3 Real Data II: Action Unit Alignment from Facial Expressions

In this experiment we aim to evaluate the performance of PCTW and DPCTW for the temporal alignment of facial expressions. Such applications can be useful for methods which require pre-aligned data, e.g. AAM. We used a portion of the MMI database [19] consisting of 66 videos of 11 different subjects. In each of these videos, a set of 3 Action Units (AUs) is activated. The videos are annotated in terms of the temporal phases of each AU (neutral, onset, apex and offset),

while the facial feature points of each subject were tracked in the same way as explained in Sec. 5.2. Thus, given a set of videos where the same set of AUs is activated by the subjects, the goal is to temporally align the phases of each AU activation across *all* videos containing that AU using the facial points. In the context of DPCTW, each $\mathbf{X}_i$ is the facial points of video i containing the same AUs, while $\mathbf{Z}|\mathbf{X}_i$ is now the common latent space given video i, the size of which is determined by cross-validation. We note that since more than one AU are activated in the same video, a perfect solution is not likely to exist, since perfectly aligning e.g. AU x may consequently lead to the misalignment of AU y.

In Fig. 6 we present results based on the number of misaligned frames for AU alignment. The facial features were perturbed with sparse spiked noise (simulating the misdetection of points with detection-based trackers) to evaluate the robustness of the proposed techniques. Values were drawn from the normal distribution $\mathcal{N}(0, 1)$ and added (uniformly) to 5% of the length of each video. We gradually increased the number of features perturbed by noise from 0 to 4. As can be seen in Fig. 6, DPCTW and PCTW outperform other methods, due to their specific noise modelling properties. The best performance is clearly obtained by DPCTW. This can be attributed to the modelling of dynamics, not only in the shared latent space of all facial point sequences but also in the domain of the individual characteristics of each sequence (in this case identifying and removing the added temporal spiked noise).

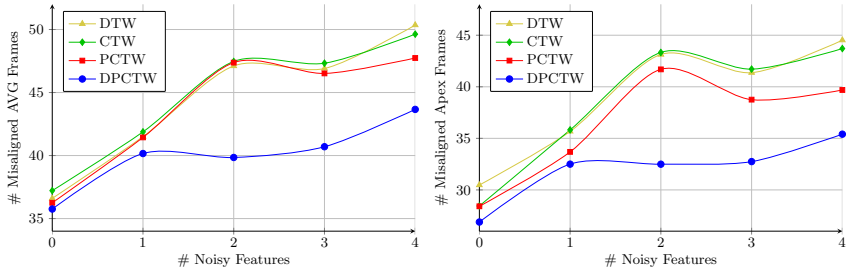

Fig. 6. Comparison of DTW, CTW, PCTW and DPCTW to the problem of action unit alignment under spiked noise added to an increasing number of features.

6 Conclusions

In this work, we presented DPCCA, a novel, dynamic & probabilistic model based on the multiset probabilistic interpretation of CCA. By integrating DPCCA with time warping, we proposed DPCTW, which can be interpreted as a unifying framework for solving the problems of (i) fusing multiple imperfect annotations and (ii) aligning temporal sequences. Our experiments show that DPCTW features such as temporal alignment, learning dynamics and identifying individual annotator/sequence factors are critical for robust performance of fusion in challenging affective behaviour analysis tasks.

7 Acknowledgements

This work is supported by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB), by the European Communitys 7th Framework Programme [FP7/2007-2013] under grant agreement no. 288235 (FROG) and by the National Science Foundation under Grant No. IIS 0916812.

References

1. Gunes, H., et al.: Emotion representation, analysis and synthesis in continuous space: A survey. In: Proc. of IEEE FG 2011 EmoSPACE WS, Santa Barbara, CA, USA (March 2011) 827–834
2. Cowie, R., McKeown, G.: Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme. http://www.semaine-project.eu/ (2010)
3. Wöllmer, M., et al.: Abandoning emotion classes. In: INTERSPEECH. (2008) 597–600
4. Nicolaou, M.A., et al.: Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Trans. on Affective Computing **2**(2) (2011) 92 –105
5. Raykar, V.C., et al.: Learning from crowds. Journal of Machine Learning Research **11** (2010) 1297–1322
6. Zhou, F., la Torre, F.D.: Canonical time warping for alignment of human behavior. In: Advances in Neural Information Processing Systems 22. (2009) 2286–2294
7. Klami, A., Kaski, S.: Probabilistic approach to detecting dependencies between data sets. Neurocomput. **72**(1-3) (2008) 39–46
8. Bach, F.R., Jordan, M.I.: A Probabilistic Interpretation of Canonical Correlation Analysis. Technical report, University of California, Berkeley (2005)
9. Zeng, Z., et al.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. PAMI. **31**(1) (2009) 39–58
10. Kim, M., Pavlovic, V.: Discriminative Learning for Dynamic State Prediction. IEEE Trans. PAMI. **31**(10) (2009) 1847–1861
11. Ghahramani, Z., Roweis, S.T.: Learning nonlinear dynamical systems using an EM algorithm. In: Advances in NIPS, MIT Press (1999) 599–605
12. Roweis, S., Ghahramani, Z.: A unifying review of linear Gaussian models. Neural Computation **11** (1999) 305–345
13. Ghahramani, Z., Jordan, M.I., Smyth, P.: Factorial hidden markov models. In: Machine Learning. Volume 29., MIT Press (1997) 245–273
14. Van der Merwe, R., Wan, E.: The square-root unscented Kalman filter for state and parameter-estimation. In: Proc. of IEEE ICASP, 2001. Volume 6. (2001) 3461–3464
15. McKeown, G., et al.: The SEMAINE corpus of emotionally coloured character interactions. In: ICME. (July 2010) 1079 –1084
16. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. United states edn. Prentice Hall (April 1993)
17. Patras, I., Pantic, M.: Particle filtering with factorized likelihoods for tracking facial features. In: Proc. of IEEE FG 2004. 97–102
18. Tipping, M.E.: Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning Research **1** (2001) 211–244
19. Pantic, M., et al.: Web-based database for facial expression analysis. In: Proc. of IEEE ICME, Amsterdam, The Netherlands (July 2005) 317–321