Supplementary material for Joint Unsupervised Deformable Spatio-Temporal Alignment of Sequences

Lazaros Zafeiriou* Epameinondas Antonakos* Stefanos Zafeiriou*‡
Maja Pantic*†
*Imperial College London †University of Twente ‡University of
*{l.zafeiriou12, e.antonakos, s.zafeiriou, m.pantic}@imperial.ac.uk, †PanticM@cs.utwente.nl, ‡

1. Experiments

The supplementary material provides additional experiments of aligning pairs of videos where the subjects perform the same AU from the MMI and UNS databases. Specifically, Fig. 1 shows the overall accuracy when aligning Brows-related AUs for each temporal phase separately. As can be seen the proposed method achieves significantly better alignment in every temporal phase compared to other methods. Moreover, it is obvious once more that by applying spatio-temporal alignment jointly (joint ARCA+DTW) we derive increased accuracy than applying spatial and temporal alignment successively (ARCA+DTW).

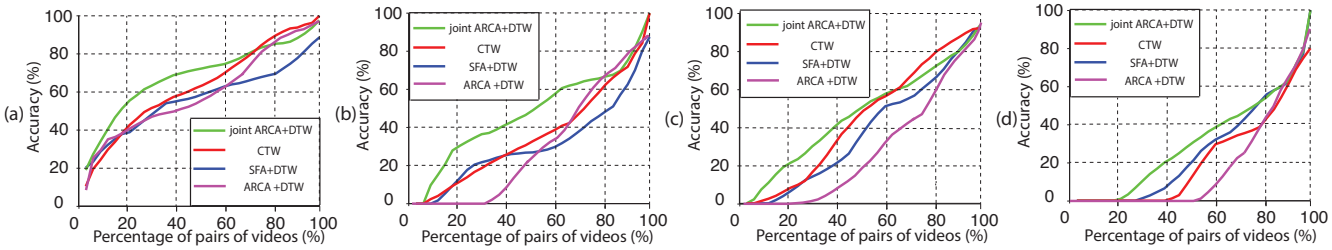


Figure 1: Overall accuracy of the Mouth-related AUs from MMI database in alignment tasks for all temporal phases in (a) Neutral phase (b) Onset phase (c) Apex phase (d) Offset phase

Fig. 2 shows the alignment accuracy when applying the proposed method in perfectly aligned images and in images where the landmark localization is achieved automatically in the UNS database. As can be seen from Fig. 2 by performing joint spatio-temporal alignment (green columns) we derive equally good results than applying the proposed method in perfectly aligned images and much worse performance when applying the proposed method with random initialization (black columns) for all temporal phases and in both spontaneous (Fig. 2 (a)) and posed (Fig. 2 (b)) smiles.

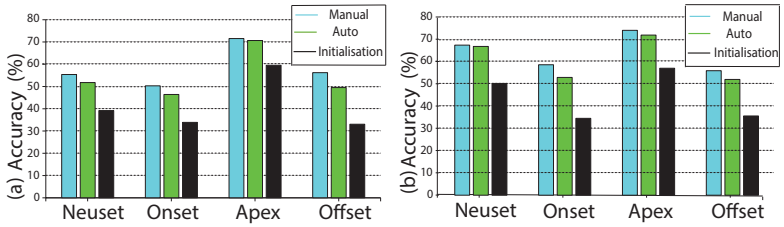
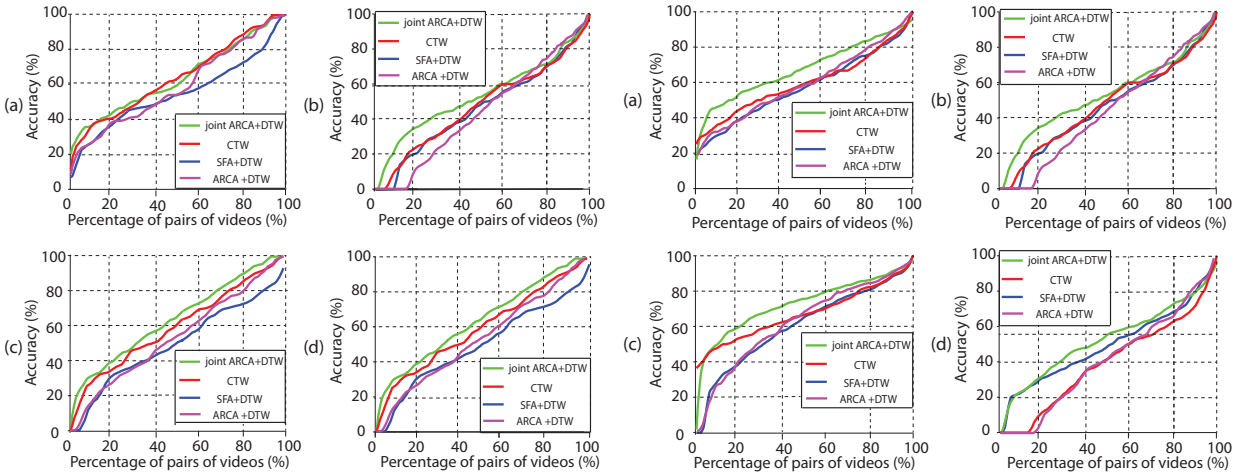


Figure 2: Performance of the proposed method applying manually annotated images (cyan columns) and with random initialisation (black columns) in UNS database for all temporal phases in (a) Spontaneous Smiles (b) Posed Smiles

Fig. 3 provides the CAD curves for the UNS database for all temporal phases in both spontaneous and posed instances. Similar to the MMI case, Figs. 3**i**, 3**ii** show the percentage of video pairs which achieved accuracy at least equal to the corresponding value



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

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

Figure 3: Temporal alignment results on UNS database.

2. Spatial Alignment Results

Herein, we evaluate the spatial alignment performance of the proposed unsupervised technique. Both MMI and UNS databases do not provide any landmarks annotations. Thus, we manually annotated the frames of 10 videos from each database with 68 landmarks, in order to provide quantitative results. The spatial alignment result is measured using a point-to-point RMS error, normalized with the face size. Table 1 reports the average initial and final errors. The initial shapes are

Initialisation	*Final shapes*
0.083	0.043

Table 1: Average point-to-point RMS error normalized with the face size on the annotated videos from MMI and UNS databases.

acquired by applying the Viola-Jones face detector on the frames of each video and fitting the mean shape $\bar{\mathbf{s}}$ in the returned bounding boxes (no random perturbations applied). The final errors are computed based on the shapes obtained at the 5th iterations of the proposed framework. The results indicate that the spatial alignment converges and the final fitted shapes are much more accurate than the initial ones.

3. Comparison with Correlated Space Component Analysis Models (Mathematical proof)

In this section, we provide the mathematical proof of deriving Eq. (15) of our manuscript which is a special case of CCA with orthogonal constraints. Recalling the Eq.(14) of our main manuscript we have

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

Assuming that both matrices of weights $\mathbf{V}_1$ and $\mathbf{V}_2$ are produced by projecting the sequences $\mathbf{X}_1$ and $\mathbf{X}_2$ onto the orthonormal bases $\mathbf{U}_1$ and $\mathbf{U}_2$, respectively (i.e., $\mathbf{V}_1 = \mathbf{U}_1^T\mathbf{X}_1$ and $\mathbf{V}_2 = \mathbf{U}_2^T\mathbf{X}_2$) we get

$$
\begin{aligned}
f(\mathbf{U}_1^o, \mathbf{U}_2^o) =& ||\mathbf{X}_1 - \mathbf{U}_1\mathbf{U}_1^T\mathbf{X}_1||_F^2 + ||\mathbf{X}_2 - \mathbf{U}_2\mathbf{U}_2^T\mathbf{X}_2||_F^2 + ||\mathbf{U}_1^T\mathbf{X}_1 - \mathbf{U}_2^T\mathbf{X}_2||_F^2 \\
=& \text{tr}\left[(\mathbf{X}_1 - \mathbf{U}_1\mathbf{U}_1^T\mathbf{X}_1)(\mathbf{X}_1 - \mathbf{U}_1\mathbf{U}_1^T\mathbf{X}_1)^T\right] + \left[(\mathbf{X}_2 - \mathbf{U}_2\mathbf{U}_2^T\mathbf{X}_2)(\mathbf{X}_2 - \mathbf{U}_2\mathbf{U}_2^T\mathbf{X}_2)^T\right] + \\
&+ \text{tr}\left[(\mathbf{U}_1^T\mathbf{X}_1 - \mathbf{U}_2^T\mathbf{X}_2)(\mathbf{U}_1^T\mathbf{X}_1 - \mathbf{U}_2^T\mathbf{X}_2)^T\right] = \text{tr}[\mathbf{X}_1\mathbf{X}_1^T] - \text{tr}[\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{U}_1] + \\
&+ \text{tr}[\mathbf{X}_2\mathbf{X}_2^T] - \text{tr}[\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{U}_2] + \text{tr}[\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{U}_1] - \text{tr}[\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_2^T\mathbf{U}_2] - \text{tr}[\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_1^T\mathbf{U}_1] + \\
&+ \text{tr}[\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{U}_2] = \text{tr}[\mathbf{X}_1\mathbf{X}_1^T] + \text{tr}[\mathbf{X}_2\mathbf{X}_2^T] - \text{tr}\left[\begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}^T \begin{pmatrix}\mathbf{0} & \mathbf{X}_1\mathbf{X}_2^T\\ \mathbf{X}_2\mathbf{X}_1^T & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}\right] \\
& \qquad \text{s.t. } \mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I},\ \mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I}
\end{aligned} \tag{2}
$$

by omitting the constants $\text{tr}[\mathbf{X}_1\mathbf{X}_1^T]$ and $\text{tr}[\mathbf{X}_2\mathbf{X}_2^T]$ we finally get

$$
\mathbf{U}_1^o, \mathbf{U}_2^o = \underset{\mathbf{U}_1, \mathbf{U}_2}{\text{argmax}}\, \text{tr}\left[\begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}^T \begin{pmatrix}\mathbf{0} & \mathbf{X}_1\mathbf{X}_2^T\\ \mathbf{X}_2\mathbf{X}_1^T & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}\right], \qquad \text{s.t.} \quad \begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}^T \begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix} = \mathbf{I}. \tag{3}
$$