Imperial College London

Joint Unsupervised Face Alignment and Behaviour Analysis

Lazaros Zafeiriou, Epameinondas Antonakos, Stefanos Zafeiriou and Maja Pantic

{l.zafeiriou12, e.antonakos, s.zafeiriou, m.pantic}@imperial.ac.uk

1. Contribution

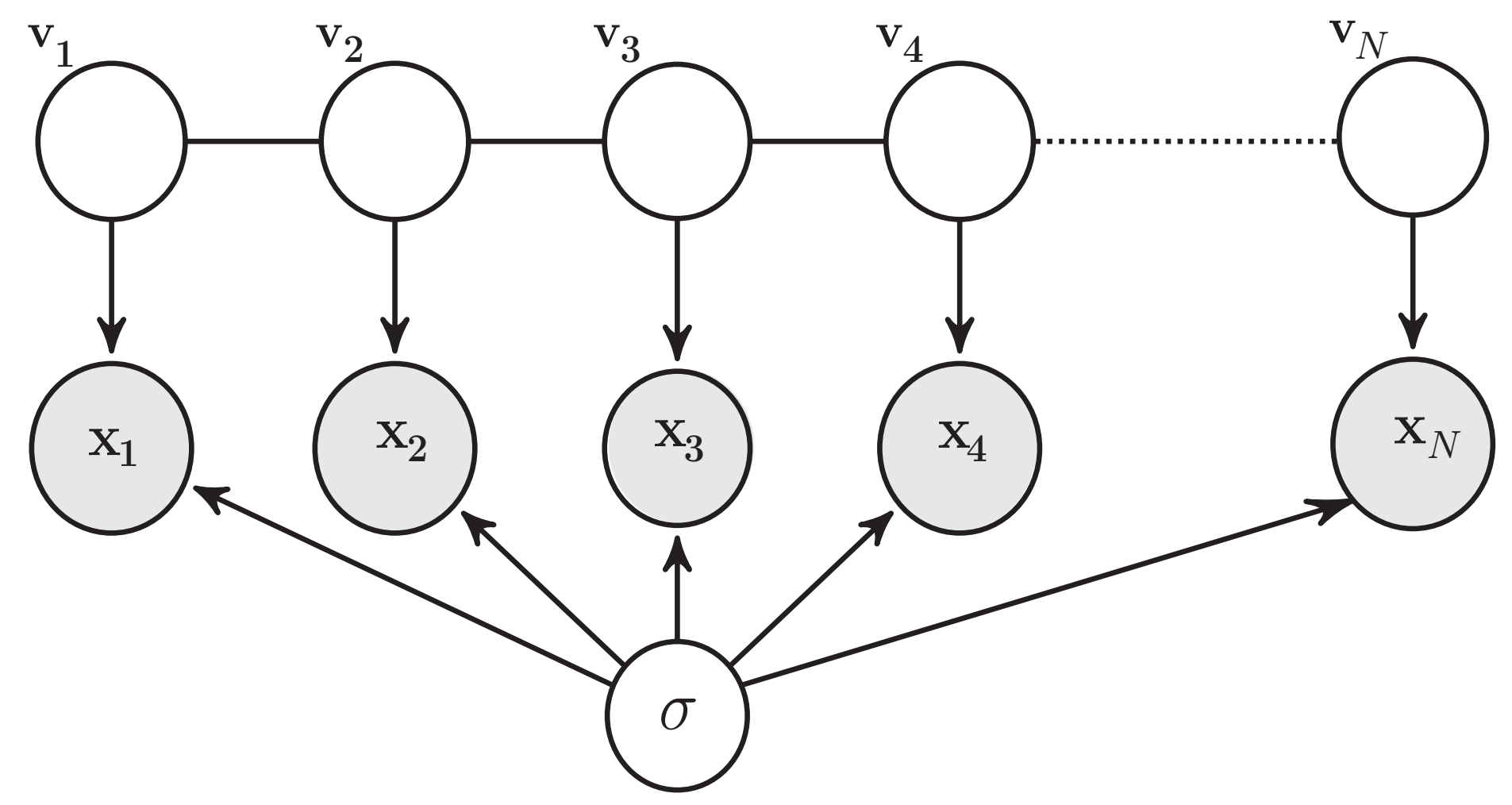
- We present a novel component analysis which can perform joint reconstruction and extraction of a latent space with first order Markov dependencies.
- We show how, by incorporating a shape model, we can perform joint alignment, i.e. facial landmarks localization, and feature extraction useful for analysis of facial events. Due to the incorporation of the motion model the extracted dynamic latent features are robust to geometric transformations.
- We show that the latent features can be used for temporal alignment of facial events.

2. Model

For deriving the ARCA we consider a probabilistic generative model which (1) captures time-variant latent features and (2) explains data generation. Assuming a generative model of the form

$$\mathbf{x}_i = \mathbf{U}\mathbf{v}_i + \mathbf{e}_i \tag{3}$$

where $\mathbf{U} \in \mathbb{R}^{F\times K}$ is a subspace of K basis, then for capturing the time-variant correlations we consider an Autoregressive model



for the latent variables $\mathbf{v}_i$ as $\mathbf{v}_i|\mathbf{v}_{i-1},\dots,\mathbf{v}_1 \sim \mathcal{N}(\mathbf{v}_i|\phi\mathbf{v}_{i-1},\mathbf{I})$ with $\mathbf{v}_1 \sim \mathcal{N}(\mathbf{v}_1|\mathbf{0},(1-\phi^2)^{-1})$.

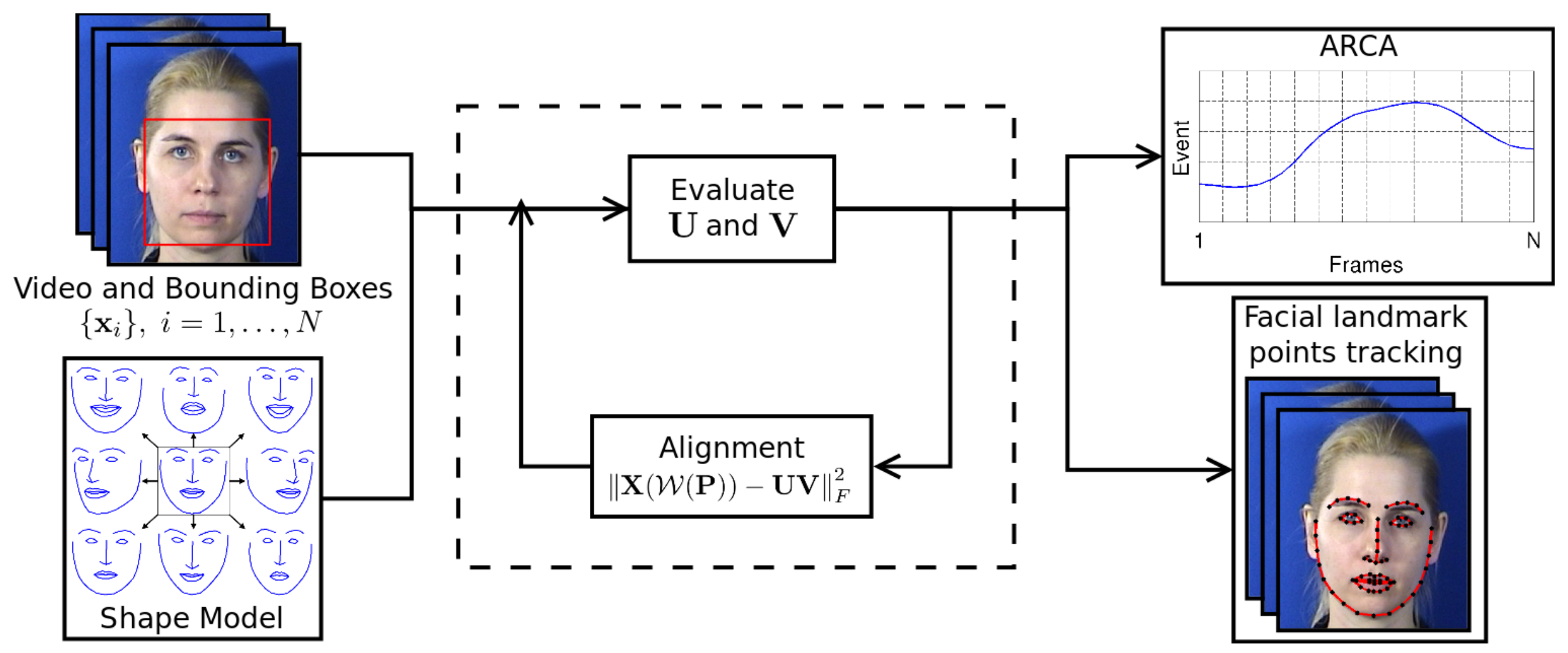
6. References

[1] Zou, Hui and Hastie, Trevor and Tibshirani, Robert. Sparse principal component analysis *Journal of computational and graphical statistics pages 265-286*

[2] Baker, Simon and Matthews, Iain. Lucas-kanade 20 years on: A unifying framework *International Journal of Computer Vision (IJVC)*

3. Method Overview

Given a video sequence with the corresponding bounding box and a shape model, the method performs jointly facial landmarks localization and spatio-temporal facial behaviour analysis


Video and Bounding Boxes
$\{\mathbf{x}_i\},\ i=1,\dots,N$
Shape Model
Evaluate
U and V
Alignment
$\|\mathbf{X}(\mathcal{W}(\mathbf{P}))-\mathbf{UV}\|_F^2$
ARCA
Event
1
N
Frames
Facial landmark points tracking

by solving the following optimization problem

$$\min_{\mathbf{U},\mathbf{V},\mathbf{P}} \quad f = \|\mathbf{X}(\mathcal{W}(\mathbf{P})) - \mathbf{UV}\|_F^2 + \lambda tr[\mathbf{VLV}^T]$$
$$\text{s.t.} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I} \tag{1}$$

The solution is given in an alternating manner consisting of two steps

Fix P and minimize with respect to $\{\mathbf{U},\mathbf{V}\}$ In this step we have a current estimate of the shape parameters matrix $\mathbf{P}$ and thus the data matrix $\mathbf{X}(\mathcal{W}(\mathbf{P}))$. In order to find the updates $\mathbf{U}$ and $\mathbf{V}$ we follow an alternative optimization framework where we fix $\mathbf{V}$ and find $\mathbf{U}$ and then fixing $\mathbf{U}$ and finding $\mathbf{V}$

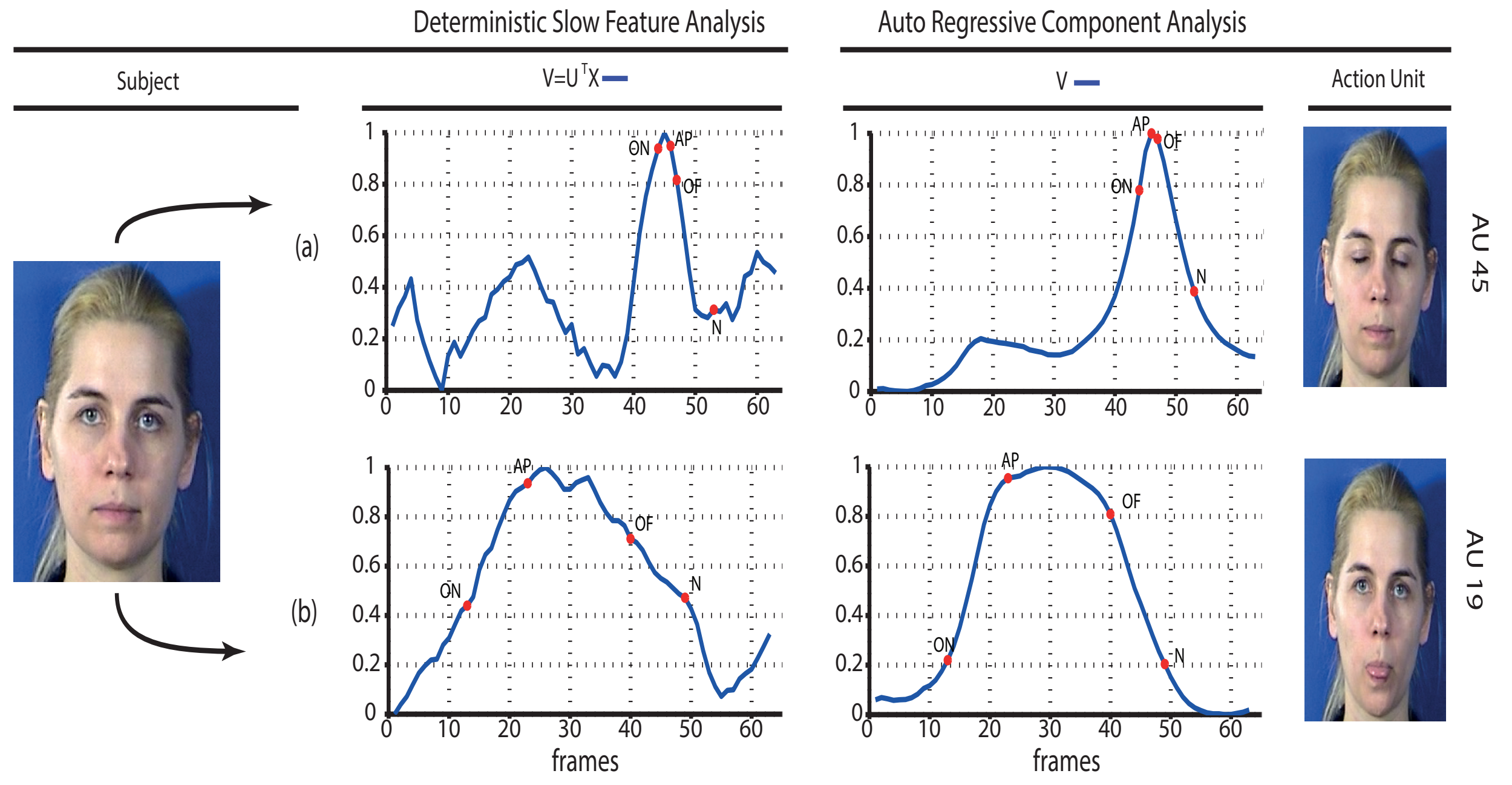
Updating U Given $\mathbf{V}$ the update for $\mathbf{U}$ is given by the skinny singular value decomposition (SSVD) of $\mathbf{X}(\mathcal{W}(\mathbf{P}))\mathbf{V}^T$ [1]

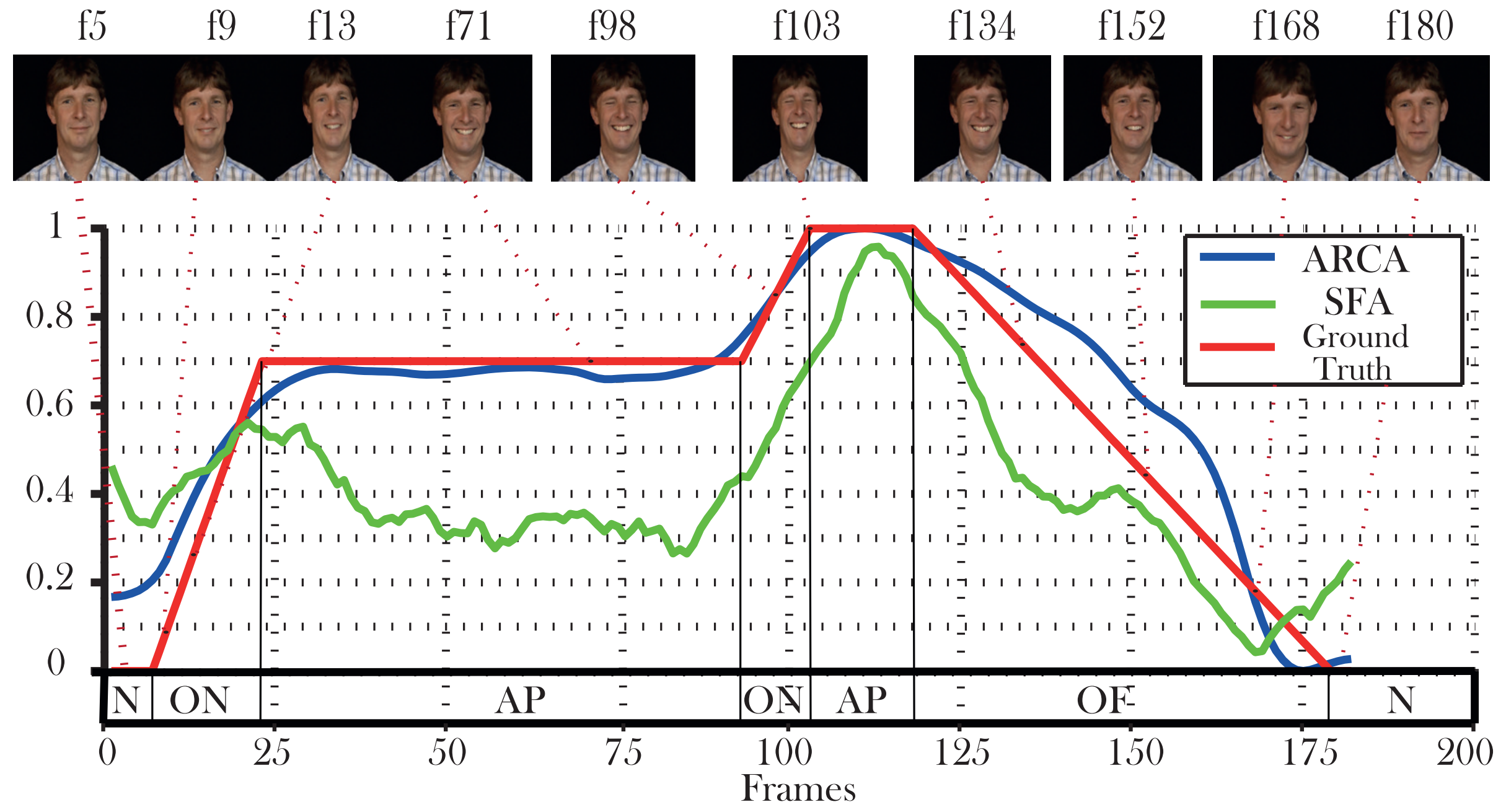
Updating V Given $\mathbf{U}$ the update for $\mathbf{V}$ is given by

$$\mathbf{V} = \mathbf{U}^T\mathbf{X}(\mathcal{W}(\mathbf{P}))(\mathbf{I}-\lambda\mathbf{L})^{-1} \tag{2}$$

Fix $\{\mathbf{U},\mathbf{V}\}$ and minimize with respect to P In this step we have a current estimation of the basis $\mathbf{U}$ and the latent features $\mathbf{V}$ and aim to estimate the motion parameters $\mathbf{P} = [\mathbf{p}_1,\dots,\mathbf{p}_N]$ for each frame, so that the Frobenius norm between the warped frames and the templates $\mathbf{UV}$ is minimized. This is achieved by using the efficient Inverse Compositional (IC) Image Alignment algorithm [2].

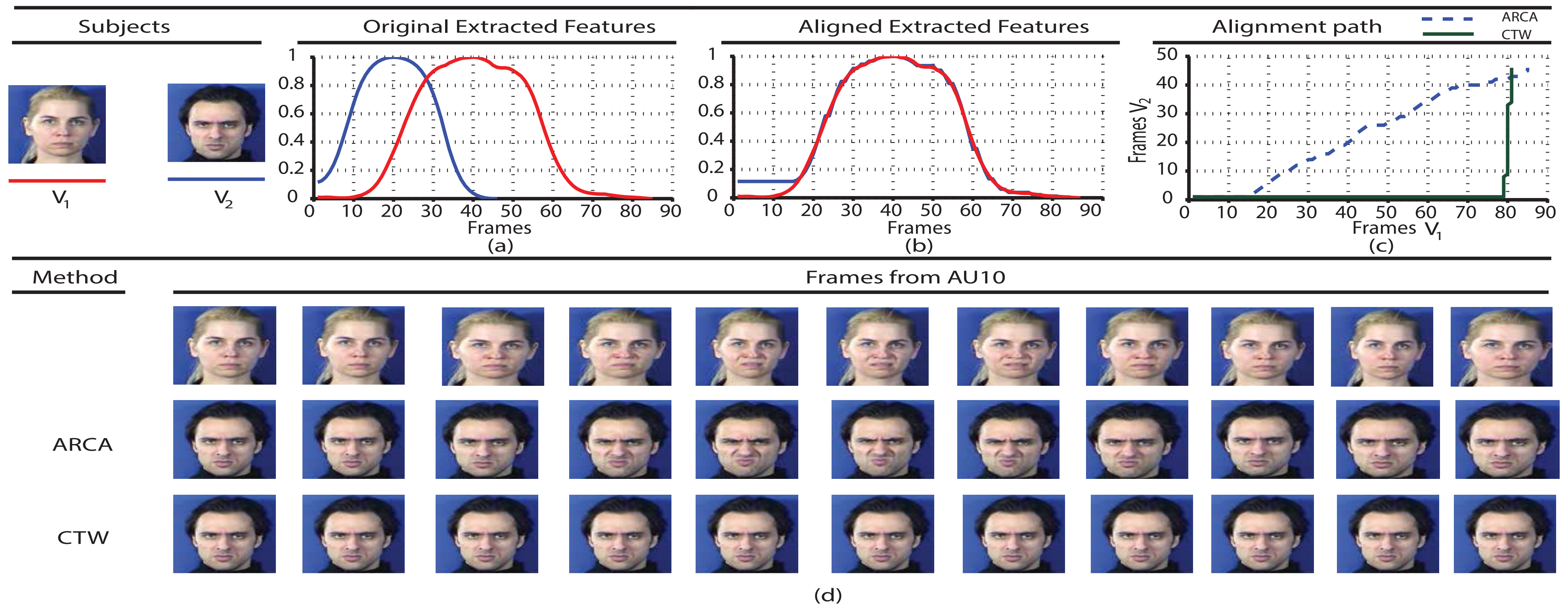
4. Results in Spatio-Temporal Behaviour Analysis





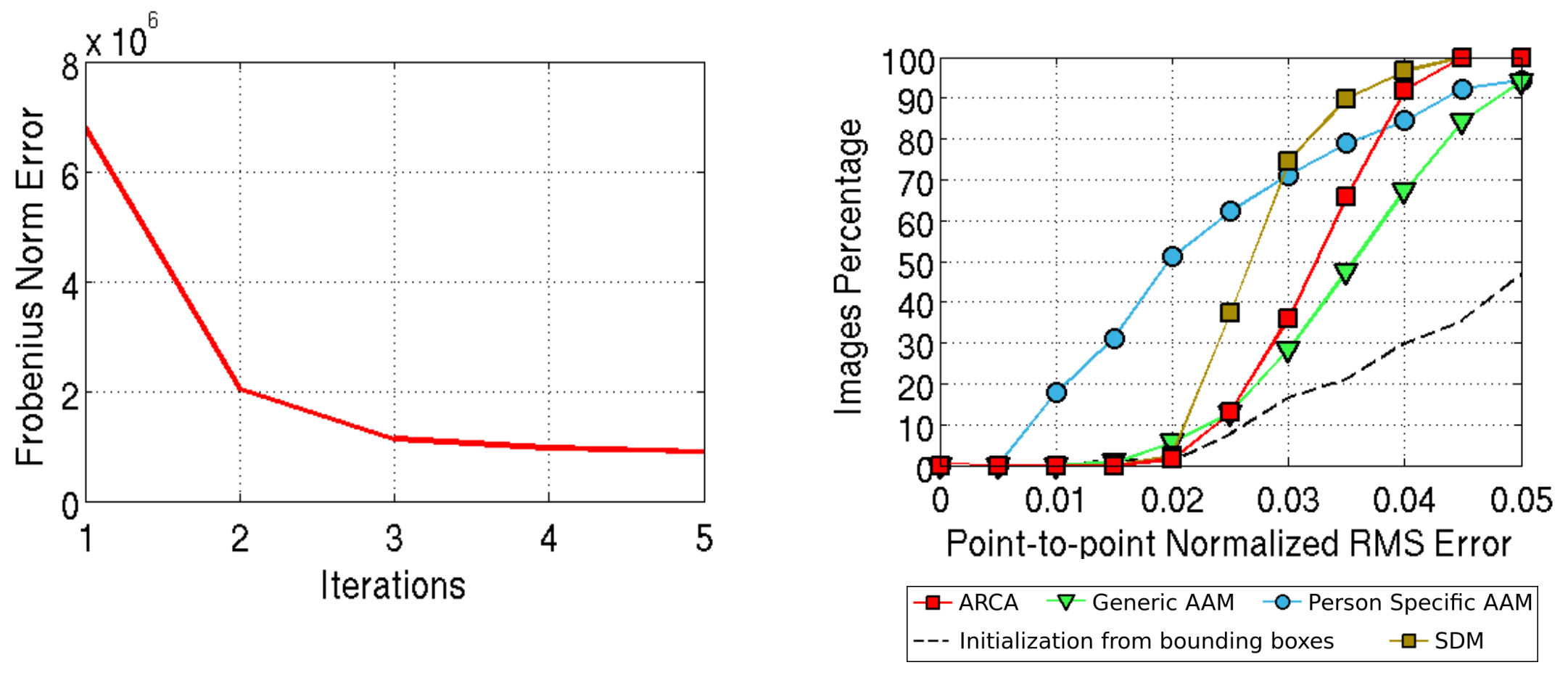
Application of SFA and ARCA on a video displaying a subject performing: (a) Blink (AU 45) and (b) Tongue Show (AU 19). The red marks indicate the ground truth moments at which the FAU's temporal phases change.

Comparison of ARCA (blue) and SFA (green) with the annotated ground truth (red) on a spontaneous video sequence from UNS database. The subject performs an FAU with multi-temporal phases (ON-Onset, AP-apex, OF-offset, N-neutral).



Aligning the AU10 performed by two different subjects. (a)Original features (b)Aligned features (c) Alignment path. (d) Frames detected form the ARCA method (second row) and CTW method (third row)

5. Results in Landmark Points Localization



Mean error over all videos per iteration and comparison of the fitting accuracy of ARCA with methods trained on manual annotations for MMI.

Indicative example of the subspace evolution on an MMI video. *Top row:* Initial subspace. *Bottom row:* Final subspace after five iterations