Nonnegative Decompositions for Dynamic Visual Data Analysis

Lazaros Zafeiriou, *Member, IEEE*, Yannis Panagakis, *Member, IEEE*, Maja Pantic, *Fellow, IEEE*, and Stefanos Zafeiriou, *Member, IEEE*

***Abstract*— The analysis of high-dimensional, possibly temporally misaligned, and time-varying visual data is a fundamental task in disciplines, such as image, vision, and behavior computing. In this paper, we focus on dynamic facial behavior analysis and in particular on the analysis of facial expressions. Distinct from the previous approaches, where sets of facial landmarks are used for face representation, raw pixel intensities are exploited for: 1) *unsupervised analysis of the temporal phases of facial expressions and facial action units (AUs)* and 2) *temporal alignment* of a certain facial behavior displayed by two different persons. To this end, the *slow features nonnegative matrix factorization* (SFNMF) is proposed in order to learn slow varying parts-based representations of time varying sequences capturing the underlying dynamics of temporal phenomena, such as facial expressions. Moreover, the SFNMF is extended in order to handle two temporally misaligned data sequences depicting the same visual phenomena. To do so, the dynamic time warping is incorporated into the SFNMF, allowing the temporal alignment of the data sets onto the subspace spanned by the estimated nonnegative shared latent features amongst the two visual sequences. Extensive experimental results in two video databases demonstrate the effectiveness of the proposed methods in: 1) unsupervised detection of the temporal phases of posed and spontaneous facial events and 2) temporal alignment of facial expressions, outperforming by a large margin the state-of-the-art methods that they are compared to.**

***Index Terms*— Nonnegative matrix factorization, slow features analysis, facial behaviour dynamics, facial expressions, temporal alignment.**

I. Introduction

THE analysis of *high-dimensional*, *dynamic* visual data arises in several vision and behaviour computing problems, where naturally occurring phenomena, such as the facial behaviour, are inherently *time-varying*. However, the high-dimensionality and the dynamic nature of such data make their modelling and analysis challenging. Indeed, the estimation and computation of models describing data with thousands of dimensions is often infeasible. To alleviate this issue, dimensionality reduction or latent feature learning methods are widely adopted [1], [2]. These methods represent the high dimensional visual data in a more compact form by means of extracted features. However, the majority of these methods neglect the temporal information and thus they cannot be applied in dynamic visual data analysis. The problem becomes more challenging when dealing with two (or multiple) high-dimensional data sequences which are also *temporally misaligned*, i.e., temporal discrepancies manifest amongst the observation sequences [3], [4].

Several dimensionality reduction methods have been proposed [1], [2]. Among them, dimensionality reduction methods which are inspired by the human visual system, has attracted significant attention in visual data analysis [5], [6]. Two prominent examples of such methods are the nonnegative matrix factorization (NMF) [5], [7], [8] and the slow feature analysis (SFA) [6]. The NMF represents nonnegative multivariate data, such as images, as a nonegative linear combination of nonnegative basis by seeking a factorization of the data matrix into two low-rank, nonnegative matrices. The nonnegativity constraint leads to interpretable parts-based representations of visual objects which is consistent to the way that the human visual cortex encodes visual information [6], [9]. The SFA is a latent feature learning method that intuitively imitates the functionality of the receptive fields of the visual cortex in time-varying stimuli [10] and hence can be exploited in analysis of dynamic visual phenomena. The temporal slowness learning principle in the SFA is motivated by the empirical observation that the semantics of sensory data, such as the objects and their attributes, are often more persistent (i.e., change smoothly) than the independent activation of any single sensory receptor. For instance, in facial behaviour analysis the SFA can learn mappings from an image sequence with rapidly varying texture to the corresponding high-level semantic concepts, that vary slowly [11], [12]. Nonetheless, the aforementioned methods cannot be applied in analysis of multiple, temporal misaligned (visual) data sequences.

A widely adopted method for temporal aliment of two data sequences is the dynamic time warping (DTW) [13]. The DTW aligns two sequences by minimizing the pairwise squared Euclidean distance via dynamic programming. Even though the DTW has been widely applied in practice, it has three main drawbacks, namely, it fails under arbitrary affine transformations of one or both sequences, it cannot handle

Manuscript received October 28, 2015; revised June 23, 2016, April 4, 2017, and July 9, 2017; accepted July 10, 2017. Date of publication August 2, 2017; date of current version September 1, 2017. This work was supported by the EPSRC Project under Grant EP/N007743/1 (FACER2VM). The work of L. Zafeiriou was supported by the EPSRC Project under Grant EP/N007743/1 (FACER2VM). The work of Y. Panagakis and M. Pantic was supported by the European Community Horizon 2020 [H2020/2014-2020] under Grant 645094 (SEWA). The work of S. Zafeiriou was supported in part by the EPSRC Project under Grant EP/J017787/1 (4D-FAB) and in part by the FiDiPro Program of Tekes under Project 1849/31/2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Scott T. Acton. *(Corresponding author: Lazaros Zafeiriou.)*

L. Zafeiriou is with AimBrain, London E14 5AB, U.K. (e-mail: lazaros.zafeiriou@gmail.com).

Y. Panagakis, M. Pantic, and S. Zafeiriou are with Imperial College London, London SW7 2AZ, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2735186

1057-7149 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

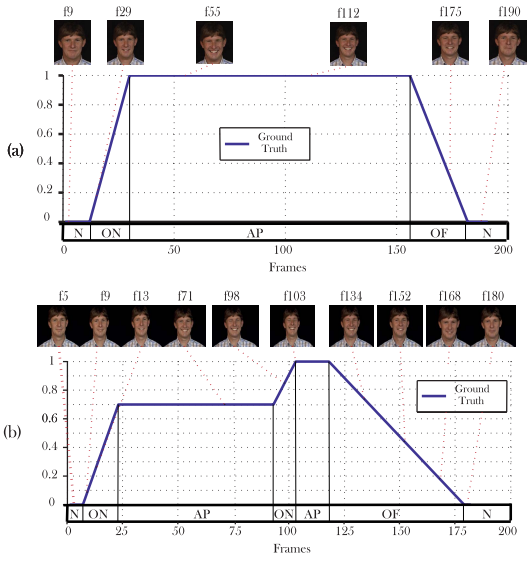

Fig. 1. Ground truth of the temporal dynamics of (a) a posed smile and (b) a spontaneous smile (N = neutral, ON = onset, AP = apex, OF = offest).

time series with different dimensions, and its performance degenerates when the data are of very high dimensions. More precisely, algorithms that rely on dynamic programming (DP), such as the the DTW, are performing suboptimal when the data are high dimensional and the computational complexity of the DP algorithms increases exponentially with dimensionality of the data. Therefore, DP algorithms are impractical for applications where high-dimensional data occur [14].

To alleviate the aforementioned drawbacks the DTW is incorporated into latent feature learning methods such as in [3], and [15]–[17]. In particular, two sequences are aligned in a shared low-dimensional latent subspace found by the canonical correlation analysis (CCA) [1], [18] or its variants [3], [16], [17]. However, the aforementioned methods ignore the temporal dynamics in time series.

In this paper, we focus on dynamic facial behaviour analysis and in particular on the analysis of facial expressions. Facial expressions encoded in terms of Facial Action Units (FAUs) activation are manifested by the motion of individual facial parts or facial muscles [19]. Therefore, facial expressions can be modelled as temporally evolving deformations of local facial parts (e.g., mouth in case of smile). The temporal dynamics of posed expressions are described by the following temporal segments: *Neutral*, *Onset*, *Apex*, and *Offset*. In particular, neutral corresponds to the phase where there is no facial motion while appex describes the temporal phase where the strongest possible facial deformation occurs. The phase where the facial motion starts until it reaches the apex is referred to as onset and the reverse path from the apex to the relaxed neutral position is the offset. Please refer to Fig. 1 (a) for a visual description of these phases. Spontaneous expressions have a different dynamical content than the posed ones exhibiting multiple apexes [20] as it is shown in Fig. 1 (b). Particularly, the movements of facial muscles in spontaneous facial expressions are smooth, synchronized, symmetrical, consistent and re ex-like, while in the posed ones the facial muscles movements are based on volitional real-time control and tend to be less smooth with more unstable dynamics [21]. For instance, it has been proved that the transitions between the temporal phases are smoother (e.g change from neutral to onset) in spontaneous compared to posed smiles. In addition, spontaneous smiles are usually accompanied by other AU/AUs and are characterised by multiple temporal phases (e.g multiple rises of the corner lips), in contrast to posed smiles. Hence, extraction and appropriate description of facial behavioural dynamics is very important for distinguishing between spontaneous and posed expressions [22]. Furthermore, recently it was shown that facial dynamics are very powerful cue towards age estimation [23].

The main idea pursued here, is to propose appropriate image decomposition methods in order to exploit raw pixel intensities for 1) *unsupervised analysis of the temporal phases of facial expressions and facial AUs* and 2) *temporal alignment* of a certain facial behaviour displayed by two different persons. To this end, two novel nonnegative matrix decompositions are proposed. The nonegativity constraints in the proposed methods are motivated by the facts that 1) pixels intensities are always non-negative and 2) the temporal activation envelope encoding the temporal phases (i.e., neutral-onset-apex-offset-neutral [22]) of the facial parts (i.e., facial muscles encoded by AUs) is always a nonnegative function of time describing the magnitude of deformations away from neutral face. Furthermore, the nonnegative muscle force constraint is used in control-based facial animation methodologies [24].

The contributions of the paper are organized as follows:

- The slow feature nonnegative matrix factorization (SFNMF) is proposed in order to learn slow varying parts-based representations of time-varying visual data depicting facial behaviour. To this end, a suitable model that combines the principles of temporal slowness and nonnegative parts-based learning is proposed in Section III. The SFNMF derives a nonnegative basis matrix capturing the activated facial part and a matrix with nonnegative coefficients representing the nonnegative latent space which accounts for the temporal activation envelope of the facial parts.
- The SFNMF is extended to handle temporally misaligned data, Section IV. To achieve this, the DTW is incorporated into the SFNMF, allowing the temporal alignment of the data sets onto the subspace spanned by the estimated nonnegative shared latent features among two visual sequences.
- Two algorithms, with guaranteed convergence to stationary point, for the SFNMF and its extension are developed in Sections III and IV, respectively.

The main advantage of the proposed methods is that the analysis of dynamic visual content and the temporal alignment do not rely on face detection, point localization, and tracking methods and therefore, they are not affected by the quality of the extracted facial landmarks. The SFNMF and its extension are evaluated in unsupervised analysis of temporal phases and in temporal warping of both posed and spontaneous facial events by conducting experiments in the MMI [25], [26] and the UvA-Nemo Smile (UNS) [27] datasets. The experimental results reported in Section V indicate that the

proposed methods outperform the methods that they are compared to.

A. Notations

Throughout the paper, matrices (vectors) are denoted by uppercase (lowercase) boldface letters e.g., $\mathbf{X}$, ($\mathbf{x}$). $\mathbf{I}$ ($\mathbf{1}$) denotes the identity matrix (vector of ones) of compatible dimensions. $\mathbf{0}$ is the zero matrix. The ith column of $\mathbf{X}$ is denoted as $\mathbf{x}_i$. The set of real numbers is denoted by $\mathbb{R}$, while the set of nonnegative real numbers is denoted by $\mathbb{R}_+$. A set of N real matrices of varying dimensions is denoted by $\{\mathbf{X}^{(n)} \in \mathbb{R}^{F_n \times T_n}\}_{n=1}^{N}$. $\|\mathbf{X}\|_F \doteq \sqrt{\sum_i \sum_j x_{ij}^2} = \sqrt{\text{tr}(\mathbf{X}^T\mathbf{X})}$ is the Frobenius norm, where tr$(\cdot)$ denotes the trace of a square matrix. The inequality $\mathbf{X} \geq \mathbf{0}$ denotes that the entries of $\mathbf{X}$ are element-wise nonnegative. The element-wise (Hadamard) product is denoted by $\circ$.

II. Background

To make the paper self-contained, this section includes a brief review of the the NMF [5], the SFA [6], the DTW [28] and the CTW [15].

A. Nonegative Matrix Factorization

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T] \in \mathbb{R}_+^{F \times T}$ be a nonnegative data matrix which containts in its collumns T, F-dimensional data points. The NMF seeks a factorization of $\mathbf{X}$ into two nonnegative, low-rank matrices $\mathbf{V} \in \mathbb{R}_+^{F \times K}$ with $K \ll F$ and $\mathbf{W} \in \mathbb{R}_+^{K \times T}$ by solving the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{V},\mathbf{W}}{\text{argmin}} \ \frac{1}{2}\|\mathbf{X} - \mathbf{V}\mathbf{W}\|_F^2 \\ &\text{s.t. } \mathbf{V} \geq \mathbf{0}, \quad \mathbf{W} \geq \mathbf{0}. \end{aligned} \tag{1}$$

$\mathbf{V}$ is the basis matrix, while $\mathbf{W}$ contains the appropriate nonegative linear combination coefficients that reconstruct each column of $\mathbf{X}$. The optimization problem (1) is solved iteratively by applying the following multiplicative update rules at each iteration, indexed by t, until a convergence criterion is met.

$$\mathbf{W}_{t+1} = \mathbf{W}_t \circ \frac{{\mathbf{V}_t}^T\mathbf{X}}{{\mathbf{V}_t}^T\mathbf{V}_t\mathbf{W}_t}, \tag{2}$$

$$\mathbf{V}_{t+1} = \mathbf{V}_t \circ \frac{\mathbf{X}{\mathbf{W}_{t+1}}^T}{\mathbf{V}_t\mathbf{W}_{t+1}{\mathbf{W}_{t+1}}^T}. \tag{3}$$

B. Slow Feature Analysis

Let us assume that $\mathbf{X} \in \mathbb{R}^{F \times T}$ represents an F-dimensional temporal sequence (e.g., T vectorized video frames). The SFA seeks a low-rank projection matrix $\mathbf{V} \in \mathbb{R}^{F \times K}$ with $K \ll F$ that extracts slowly varying features from the rapid varying input sequence $\mathbf{X}$ by solving the following optimization problem:

$$\underset{\mathbf{V}}{\text{argmin}} \ \text{tr}[\mathbf{V}^T\mathbf{A}\mathbf{V}], \quad \text{s.t. } \mathbf{V}^T\mathbf{B}\mathbf{V} = \mathbf{I}. \tag{4}$$

In (4), $\mathbf{A}$ is the covariance matrix of the first-order temporal derivative of $\mathbf{X}$, denoted as $\dot{\mathbf{X}}$, and $\mathbf{B}$ is the data covariance matrix. That is,

$$\mathbf{A} = \frac{1}{T-1}\dot{\mathbf{X}}\dot{\mathbf{X}}^T = \frac{1}{T-1}\mathbf{X}\mathbf{L}\mathbf{X}^T, \quad \mathbf{B} = \frac{1}{T}\mathbf{X}\mathbf{X}^T, \tag{5}$$

where $\mathbf{L} = \mathbf{P}\mathbf{P}^T$ and $\mathbf{P}$ is an $T \times (T-1)$ matrix with elements $p_{i,i} = -1$ and $p_{i+1,i} = 1$. The solution of (4) is found by the Generalized Eigenvalue Problem $\mathbf{A}\mathbf{V} = \mathbf{B}\mathbf{V}\Lambda$, where the columns of the projection matrix $\mathbf{V}$ are the generalized eigenvectors associated with the K lowest eigenvalues contained in the diagonal matrix Λ [6].

C. Dynamic Time Warping

Given two temporally misaligned data sets $\{\mathbf{X}^{(n)} = [\mathbf{x}_1^{(n)}, \mathbf{x}_2^{(n)}, \cdots, \mathbf{x}_{T_n}^{(n)}] \in \mathbb{R}^{F \times T_n}\}_{n=1}^2$, with $T_1 \neq T_2$ the DTW aligns them along the time axis by solving [28]:

$$\underset{\{\Delta^{(n)}\}_{n=1}^2}{\text{argmin}} \ \frac{1}{2}\|\mathbf{X}^{(1)}\Delta^{(1)} - \mathbf{X}^{(2)}\Delta^{(2)}\|_F^2, \tag{6}$$

where $\Delta^{(1)} = \Delta(\mathbf{p}^{(1)}) \in \{0,1\}^{T_1 \times m}$ and $\Delta^{(2)} = \Delta(\mathbf{p}^{(2)}) \in \{0,1\}^{T_2 \times m}$ are binary selection matrices associated with the warping paths ($\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$) by a non-linear mapping, $\Delta(\mathbf{p}) : \{1:T\}^m \rightarrow \{0,1\}^{T \times m}$, which sets $\delta_{p_t,t} = 1$ for $t \in \{1:m\}$ and zero otherwise, where $m \geq max(T_1, T_2)$ is the number of steps required to align both time series and is optimally selected by the DTW algorithm. The warping paths $\mathbf{p}^{(1)} \in \{1:T_1\}^m$ and $\mathbf{p}^{(2)} \in \{1:T_2\}^m$ indicate the compound of alignment in frames. For instance, the i^{th} frame in $\mathbf{X}^{(1)}$ and the j^{th} frame in $\mathbf{X}^{(2)}$ are aligned if there exists $p_t^{(1)} = i$ and $p_t^{(2)} = j$ for some t.

Additionally, in order the time series to be aligned in time, the warping paths has to satisfy the boundary condition ($[p_1^{(1)}, p_1^{(2)}] \equiv [1,1]^T$ and $[p_m^{(1)}, p_m^{(2)}] \equiv [T_1, T_2]$), the continuity condition $[p_t^{(1)}, p_t^{(2)}] - [p_{t-1}^{(1)}, p_{t-1}^{(2)}] \in \{[0,1],[1,0],[1,0]\}$.) and the monotonicity condition ($t_1 \geq t_2 \Rightarrow p_{t_1}^{(1)} \geq p_{t_2}^{(1)}$ and $p_{t_1}^{(2)} \geq p_{t_2}^{(2)}$).

Although the number of possible alignments is exponential in $T_1 \cdot T_2$, the DTW recovers the optimal alignment path in $\mathcal{O}(T_1 \cdot T_2)$ by employing dynamic programming. Clearly, the DTW can handle only data of the same dimensions.

D. Canonical Time Warping

The CTW [15] incorporates CCA [1] into the DTW, allowing the alignment of data sequences of different dimensions by projecting them into a common latent subspace found by the CCA. Furthermore, the CCA-based projections perform feature selection by reducing the dimensionality of the data to that of the common latent subspace, handling the irrelevant or possibly noisy attributes.

More formally, let $\{\mathbf{X}^{(n)} \in \mathbb{R}^{F_n \times T_n}\}_{n=1}^2$ be a set of temporally misaligned data of different dimensionality (i.e., $F_1 \neq F_2$), the CCA is incorporated into the DTW by solving [15]:

$$\begin{aligned} &\underset{\{\mathbf{V}^{(n)},\Delta^{(n)}\}_{n=1}^2}{\text{argmin}} \ \frac{1}{2}\|{\mathbf{V}^{(1)}}^T\mathbf{X}^{(1)}\Delta^{(1)} - {\mathbf{V}^{(2)}}^T\mathbf{X}^{(2)}\Delta^{(2)}\|_F^2, \\ &\text{s.t. } {\mathbf{V}^{(n)}}^T\mathbf{X}^{(n)}{\mathbf{X}^{(n)}}^T\mathbf{V}^{(n)} = \mathbf{I}, \\ &\quad {\mathbf{V}^{(1)}}^T\mathbf{X}^{(1)}\Delta^{(1)}{\Delta^{(2)}}^T{\mathbf{X}^{(2)}}^T\mathbf{V}^{(2)} = \mathbf{D}, \\ &\quad \mathbf{X}^{(n)}\Delta^{(n)}\mathbf{1} = \mathbf{0}, \quad \Delta^{(n)} \in \{0,1\}^{J_n \times J}, \quad n = 1,2. \end{aligned} \tag{7}$$

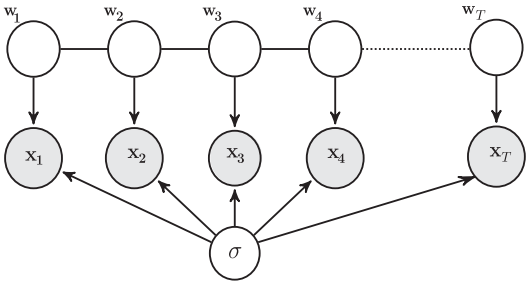

Fig. 2. Graphical model of an Autoregressive process.

$\mathbf{V}^{(1)} \in \mathbb{R}^{F_1 \times K}$ and $\mathbf{V}^{(2)} \in \mathbb{R}^{F_2 \times K}$ project $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively onto a common latent subspace of $K \leq \min(F_1, F_2)$ dimensions, where the correlation between the data sequences is maximized. $\mathbf{D}$ is a diagonal matrix of compatible dimensions. The set of constraints in (7) is imposed in order to make the CTW translation, rotation, and scaling invariant. The solution of (7) is obtained by solving CCA and DTW in an alternating fashion.

III. Slow Features Nonnegative Matrix Factorization

In this section, the SFNMF is detailed. In particular, the optimization problem of the SFNMF is derived from a probabilistic point of view by introducing an autoregressive statistical model for capturing temporal dependencies (Section III-A). An iterative algorithm for the SFNMF is proposed in Section III-C.

A. Autoregressive Model for Capturing Temporal Dependencies

Let $\mathbf{X} \in \mathbb{R}^{F \times T}$ represents a time-variant, high-dimensional time-series e.g., a video sequence of T frames depicting a person performing a facial expression. We assume that the columns of $\mathbf{X}$ are described by the following autoregressive (AR) model:

$$\begin{aligned} \mathbf{x}_i &= \mathbf{V}\mathbf{w}_i + \mathbf{e}_i, \quad \mathbf{e}_i \sim \mathcal{N}(\mathbf{e}_i|\mathbf{0}, \sigma^2\mathbf{I}) \\ \mathbf{w}_i &= \phi\mathbf{w}_{i-1} + \mathbf{n}_i, \quad \mathbf{n}_i \sim \mathcal{N}(\mathbf{n}_i|\mathbf{0}, \mathbf{I}) \\ \mathbf{w}_i &\sim \mathcal{N}(\mathbf{w}_i|\mathbf{0}, (1-\phi^2)^{-1}), \end{aligned} \tag{8}$$

where $\mathbf{V} \in \mathbb{R}^{F \times K}$ is a linear subspace of K basis ($K < \min(F, T)$), $\mathbf{w}_i \in \mathbb{R}^K$ are the latent features, and ϕ are coefficient regulating the first order dependencies between successive latent variables. The graphical model for such an AR model is depicted in Fig. 2.

Let the latent features stored in columns of $\mathbf{W} \in \mathbb{R}^{K \times T}$ and $\tilde{\mathbf{w}}_j \in \mathbb{R}^{T \times 1}$ be the j-th row of $\mathbf{W}$, the prior over the latent variables is assumed to be:

$$p(\tilde{\mathbf{w}}_j|\mathbf{L}) = \frac{|\mathbf{L}|}{\sqrt{(2\pi)^K}} e^{-\frac{1}{2}(\tilde{\mathbf{w}}_j)^T \mathbf{L} \tilde{\mathbf{w}}_j}. \tag{9}$$

Since the autoregressive model (8) is a special case of a Gaussian Markov Random Field (GMRF) [29], $\mathbf{L} \in \mathbb{R}^{T \times T}$ is a tri-diagonal precision matrix defined as follows:

$$\mathbf{L} = \begin{pmatrix} 1 & -\phi & & & \\ -\phi & 1+\phi^2 & -\phi & & \\ & \ddots & \ddots & \ddots & \\ & & -\phi & 1+\phi^2 & -\phi \\ & & & -\phi & 1 \end{pmatrix} \tag{10}$$

Therefore, the prior for all the rows of matrix $\mathbf{W}$ is written as

$$\begin{aligned} p(\mathbf{W}|\mathbf{L}) &= \prod_{j=1}^{K} p(\tilde{\mathbf{w}}_j|\mathbf{L}) = \frac{|\mathbf{L}|^T}{\sqrt{(2\pi)^{KT}}} e^{-\frac{1}{2}\sum_{j=1}^{T}(\tilde{\mathbf{w}}_j)^T \mathbf{L} \tilde{\mathbf{w}}_j} \\ &= \frac{|\mathbf{L}|^T}{\sqrt{(2\pi)^{KT}}} e^{-\frac{1}{2}\mathrm{tr}[\mathbf{W}\mathbf{L}\mathbf{W}^T]}. \end{aligned} \tag{11}$$

Hence, according to Fig. 2, the factorization of the joint likelihood of $\mathbf{X}, \mathbf{W}$ given σ^2, $\mathbf{L}$ and $\mathbf{V}$ has the form

$$\begin{aligned} &p(\mathbf{X}, \mathbf{W}|\mathbf{L}, \mathbf{V}, \sigma^2) \\ &= p(\mathbf{X}|\mathbf{W}, \mathbf{V}, \sigma^2) p(\mathbf{W}|\mathbf{L}) \\ &= \prod_{i=1}^{T} p(\tilde{\mathbf{x}}_i|\mathbf{w}_i, \mathbf{L}, \sigma^2) p(\mathbf{W}|\mathbf{L}) \\ &= \frac{|\mathbf{L}|^T}{\sqrt{(\sigma^2)^{FT}(2\pi)^{T(K+F)}}} e^{-\frac{1}{2}\left(\frac{1}{\sigma^2}\|\mathbf{X}-\mathbf{V}\mathbf{W}\|_F^2 + \mathrm{tr}[\mathbf{W}\mathbf{L}\mathbf{W}^T]\right)} \end{aligned} \tag{12}$$

It is easy to show that the Maximum Likelihood (ML) solution of (12) in the deterministic case is given by solving:

$$\underset{\mathbf{V},\mathbf{W}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{V}\mathbf{W}\|_F^2 + \lambda \mathrm{tr}[\mathbf{W}\mathbf{L}\mathbf{W}^T], \tag{13}$$

where $\lambda \geq 0$ is a regularization parameter balancing the two terms in (13). In particular, $\|\mathbf{X}-\mathbf{V}\mathbf{W}\|_F^2$ measures how well the data can be reconstructed by the product of the basis matrix $\mathbf{V}$ and the latent space weights $\mathbf{W}$, while the second term $\mathrm{tr}[\mathbf{W}\mathbf{L}\mathbf{W}^T]$ models the undirected temporal dependencies.

B. SFNMF Optimization Problem

Although, the the temporal dependencies in data are explicitly modelled in (13), its solution does not explain the data as purely additive linear combination of nonnegative basis which is desirable in case of visual data analysis. To alleviate this issue, the SFNMF imposes nonnegativity constraints in (13) by solving the non-linear optimization problem:

$$\begin{aligned} &\underset{\mathbf{V},\mathbf{W}}{\operatorname{argmin}} \;\; F(\mathbf{V}, \mathbf{W}) = \|\mathbf{X} - \mathbf{V}\mathbf{W}\|_F^2 + \lambda \mathrm{tr}[\mathbf{W}\mathbf{L}\mathbf{W}^T] \\ &\text{s.t. } \mathbf{V} \geq \mathbf{0}, \quad \mathbf{W} \geq \mathbf{0}. \end{aligned} \tag{14}$$

In (14), $\mathbf{V} \in \mathbb{R}_+^{F \times K}$ are the nonnegative basis matrix accounting for the active facial parts and $\mathbf{W} \in \mathbb{R}_+^{K \times T_n}$ are the coefficient matrices capturing the dynamics of the facial event (i.e., temporal envelope).

C. Multiplicative Update Rules for SFNMF Optimization

To solve the SFNMF constrained optimization problem in (14) a block-coordinate descent procedure is employed, where $\mathbf{V}$ and $\mathbf{W}$ are updated iteratively via the multiplicative updates derived next. Let us introduce the Lagrangian

multipliers $\Phi \in \mathbb{R}^{F\times K}$ and $\Psi \in R^{K\times T}$ associated with the inequality constraints. Thus, the Lagrangian function $\mathcal{L}(\mathbf{V},\mathbf{W})$ is expressed as:

$$\mathcal{L}(\mathbf{V},\mathbf{W}) = \|\mathbf{X}-\mathbf{V}\mathbf{W}\|_F^2 + \lambda\ \mathrm{tr}[\mathbf{W}\mathbf{L}\mathbf{W}^T] + \mathrm{tr}[\mathbf{\Phi}\mathbf{V}^T] + \mathrm{tr}[\mathbf{\Psi}\mathbf{W}^T]. \quad (15)$$

To derive multiplicative updates, we set partial derivatives of the Lagrangian function with respect to $\mathbf{V}^{(n)}$ and $\mathbf{W}^{(n)}$ equal to zero. Let $\mathbf{L}^{(n)}$ be decomposed into two nonnegative parts i.e., $\mathbf{L}^{(n)} = \mathbf{L}^{(n)^+} - \mathbf{L}^{(n)^-}$ as follows.

$$\mathbf{L}^{(n)^+} = \frac{(|\mathbf{L}^{(n)}| + \mathbf{L}^{(n)})}{2} \quad (16)$$

$$\mathbf{L}^{(n)^-} = \frac{(|\mathbf{L}^{(n)}| - \mathbf{L}^{(n)})}{2} \quad (17)$$

The partial derivatives are given by,

$$\frac{\partial \mathcal{L}}{\partial v_{i,k}} = -2[\mathbf{X}\mathbf{W}^T]_{i,k} + 2[\mathbf{V}\mathbf{W}\mathbf{W}^T]_{i,k} + \phi_{i,k} = 0 \quad (18)$$

$$\frac{\partial \mathcal{L}}{\partial w_{k,j}} = 2[\mathbf{V}^T\mathbf{V}\mathbf{W}]_{k,j} - 2[\mathbf{V}^T\mathbf{X}]_{k,j} + 2\lambda[\mathbf{W}\mathbf{L}^+]_{k,j} - 2\lambda[\mathbf{W}\mathbf{L}^-]_{k,j} + \psi_{k,j} = 0. \quad (19)$$

Let t be the iteration index. By employing the Karush-Kuhn-Tucker conditions $\phi_{i,k}v_{i,k} = 0$ and $\psi_{k,j}w_{k,j} = 0$ the following multiplicative updates are derived:

$$\mathbf{V}_{t+1} = \mathbf{V}_t \circ \frac{\mathbf{X}\mathbf{W}_{t+1}{}^T}{\mathbf{V}_t\mathbf{W}_{t+1}\mathbf{W}_{t+1}{}^T}. \quad (20)$$

$$\mathbf{W}_{t+1} = \mathbf{W}_t \circ \frac{\mathbf{V}_t{}^T\mathbf{X} + \lambda\mathbf{W}_t\mathbf{L}^-}{\mathbf{V}_t{}^T\mathbf{V}_t\mathbf{W}_t + \lambda\mathbf{W}_t\mathbf{L}^+}. \quad (21)$$

The main limitation of the above multiplicative updates is that they do no guarantee convergence to stationary point [30]. To alleviate this, modified multiplicative update rules are developed next.

D. Modified Multiplicative Updates

The multiplicative updates (20) and (21) are equivalently written in a gradient descent form:

$$\mathbf{V}_{t+1} = \mathbf{V}_t - \mathbf{V}_t \circ \frac{\nabla_V F(\mathbf{V}_t,\mathbf{W}_{t+1})}{2\mathbf{V}_t\mathbf{W}_{t+1}\mathbf{W}_{t+1}{}^T}, \quad (22)$$

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mathbf{W}_t \circ \frac{\nabla_W F(\mathbf{V}_t,\mathbf{W}_t)}{2(\mathbf{V}_t{}^T\mathbf{V}_t\mathbf{W}_t + \lambda\mathbf{W}_t\mathbf{L}^+)}, \quad (23)$$

with

$$\frac{\mathbf{V}_t}{2\mathbf{V}_t\mathbf{W}_t\mathbf{W}_t{}^T}, \quad \frac{\mathbf{W}_t}{2(\mathbf{V}_t{}^T\mathbf{V}_t\mathbf{W}_t + \lambda\mathbf{W}_t\mathbf{L}^+)}$$

being the step sizes when updating $\mathbf{V}_t$ and $\mathbf{W}_t$, respectively.

Unfortunately, by applying the above update rules we cannot guarantee that the SFNMF objective function is strictly decreasing due to the following reasons: 1) the step sizes may have zero denominator, and 2) if the nominator of the step sizes is zero and the gradient $\nabla_V F(\mathbf{V}_t,\mathbf{W}_t) < 0$ ($\nabla_W F(\mathbf{V}_t,\mathbf{W}_t) < 0$) then $\mathbf{V}_t$ ($\mathbf{W}_t$) does not change. Therefore we cannot ensure convergence to stationary point [30].

In order to overcome the aforementioned limitations we follow [30] and modify the step sizes as follows:

$$\frac{\overline{\mathbf{V}}_t}{2\mathbf{V}_t\mathbf{W}_t\mathbf{W}_t{}^T}, \quad \frac{\overline{\mathbf{W}}_t}{2(\mathbf{V}_t{}^T\mathbf{V}_t\overline{\mathbf{W}}_t + \lambda\overline{\mathbf{W}}_t\mathbf{L}^+) + \delta} \quad (24)$$

where

$$\overline{\mathbf{V}}_t \equiv \begin{cases} \mathbf{V}_t, & if\ \nabla_V F(\mathbf{V}_t,\mathbf{W}_t) \geq 0 \\ \max(\mathbf{V}_t,\sigma), & if\ \nabla_V F(\mathbf{V}_t,\mathbf{W}_t) < 0 \end{cases} \quad (25)$$

$$\overline{\mathbf{W}}_t \equiv \begin{cases} \mathbf{W}_t, & if\ \nabla_W F(\mathbf{V}_t,\mathbf{W}_t) \geq 0 \\ \max(\mathbf{W}_t,\sigma), & if\ \nabla_W F(\mathbf{V}_t,\mathbf{W}_t) < 0. \end{cases} \quad (26)$$

δ and σ represent small positive numbers. The modified update rules for the SFNMF is summarized in Algorithm 1 where its convergence is given in the Appendix A.

Algorithm 1 SFNMF

Data: $\mathbf{X} \in \mathbb{R}_+^{F\times T}, \mathbf{L} \in \mathbb{R}^{T\times T}, 1 \leq K \leq \min\{F,T\}$
Result: $\mathbf{V} \in \mathbb{R}_+^{F\times K}, \mathbf{W} \in \mathbb{R}_+^{K\times T}$
Give $\delta > 0$ and $\sigma > 0$. Initialize: $\mathbf{V}_1 \geq \mathbf{0}, \mathbf{W}_1 \geq \mathbf{0}$, $t = 1$
while *($\mathbf{V}_t$,$\mathbf{W}_t$) not a stationary point* **do**
  Update $\mathbf{W}$ according to

$$\mathbf{W}_{t,r} = \mathbf{W}_t - \frac{\overline{\mathbf{W}}_t \circ \nabla_W F(\mathbf{V}_t,\mathbf{W}_t)}{2(\mathbf{V}_t{}^T\mathbf{V}_t\overline{\mathbf{W}}_t + \lambda\overline{\mathbf{W}}_t\mathbf{L}^+) + \delta} \quad (27)$$

  Update $\mathbf{V}$ according to

$$\mathbf{V}_{t,r} = \mathbf{V}_t - \frac{\overline{\mathbf{V}}_t \circ \nabla_V F(\mathbf{V}_t,\mathbf{W}_{t,r})}{2\overline{\mathbf{V}}_t\mathbf{W}_{t,r}\mathbf{W}_{t,r}{}^T + \delta} \quad (28)$$

  Normalise $\mathbf{W}_{t,r}$ and $\mathbf{V}_{t,r}$ to $\mathbf{V}_t$ and $\mathbf{W}_t$ respectively, so that the sum of $\mathbf{V}_t$ columns is one. In case where the whole column is zero, then this and the corresponding row in $\mathbf{V}$ do not change.
  $t \leftarrow t+1$

E. Computational Complexity

The computational complexity of the SFNMF is as follows. The cost of calculating the update rules (22) and (23) is identical to that of NMF, namely $\mathcal{O}(tFTK)$ with t being the total number of iterations. Apart from the multiplicative updates, SFNMF also needs to construct the precision matrix $\mathbf{L}$ which takes $\mathcal{O}(T^2F)$ making the overall cost for the SFNMF to be $\mathcal{O}(tFTK + T^2F)$

Similarly to the SFNMF, the GNMF requires $\mathcal{O}(T^2F)$ operations to construct the p-nearest neighbor graph, and thus its overall cost is identical with that of the SFNMF.

IV. SFNMF With Time Warping

Accurate temporal alignment of nonnegative data sequences is an essential pre-processing step towards the analysis of multiple, temporally misaligned data sequences depicting the same visual phenomena. The problem is defined as finding the temporal coordinate transformation that brings two given data sequences into alignment in time. To handle temporally misaligned, nonnegative data sequences, the DTW is

incorporated into the SFNMF. The proposed method is coined as SFNMF-TW. Formally, given two data sequences depicting the same facial event, $\{\mathbf{X}^{(n)} \in \mathbb{R}_+^{F_n \times T_n}\}_{n=1}^2$, of different dimensionality and length, i.e., $F_1 \neq F_2, T_1 \neq T_2$, the SFNMF-TW enables their temporal alignment onto the subspace spanned by the estimated shared latent features.

To this end, the SFNMF-TW solves:

$$\begin{aligned}\min_{\{\mathbf{V}^{(n)},\mathbf{W}^{(n)}\Delta^{(n)}\}_{n=1}^2} & \sum_{n=1}^{2} \|(\mathbf{X}^{(n)} - \mathbf{V}^{(n)}\mathbf{W}^{(n)})\Delta^{(n)}\|_F^2 \\ & + \sum_{n=1}^{2} \lambda \mathrm{tr}[\mathbf{W}^{(n)}\mathbf{L}^{(n)}\mathbf{W}^{(n)^T}] \\ & + \|\mathbf{W}^{(1)}\Delta^{(1)} - \mathbf{W}^{(2)}\Delta^{(2)}\|_F^2 \\ \text{s.t. } \{\mathbf{V}^{(n)} \geq \mathbf{0}, \mathbf{W}^{(n)} \geq \mathbf{0}, & \Delta^{(n)} \in \{0,1\}^{T_n \times T}\}_{n=1}^2, \quad (29)\end{aligned}$$

where, $\mathbf{V}^{(n)} \in \mathbb{R}_+^{F_n \times K}$ are the nonnegative basis matrices accounting for the active facial parts and $\mathbf{W}^{(n)} \in \mathbb{R}_+^{K \times T_n}$ are the coefficient matrices capturing the temporal dynamics of the facial event. $\{\mathbf{L}^{(n)}\}_{n=1}^2$ are tri-diagonal precision matrices of the form defined in (10) and $\{\Delta^{(n)}\}_{n=1}^2$ are binary selection matrices encoding the alignment path as in the DTW.

Again, (29) is solved using a block-coordinate descent procedure, where the matrices $\{\mathbf{V}^{(n)}, \mathbf{W}^{(n)}\}_{n=1}^2$ are updated via multiplicative update rules and the warping paths via the DTW at each iteration. Specifically, to solve (29) we introduce the Lagrangian multipliers $\{\mathbf{\Phi}^{(n)} \in \mathbb{R}^{F_n \times K}\}_{n=1}^2$ and $\{\mathbf{\Psi}^{(n)} \in \mathbb{R}^{K \times T_n}\}_{n=1}^2$, associated with the inequality constraints. The Lagrangian function for (29) is formulated as:

$$\begin{aligned}\mathcal{L}(\{\mathbf{V}^{(n)}, \mathbf{W}^{(n)}, \Delta^{(n)}\}_{n=1}^2) &= \sum_{n=1}^{2} \|(\mathbf{X}^{(n)} - \mathbf{V}^{(n)}\mathbf{W}^{(n)})\Delta^{(n)}\|_F^2 \\ &+ \sum_{n=1}^{2} \lambda \mathrm{tr}[\mathbf{W}^{(n)}\mathbf{L}^{(n)}\mathbf{W}^{(n)^T}] + \|\mathbf{W}^{(1)}\Delta^{(1)} - \mathbf{W}^{(2)}\Delta^{(2)}\|_F^2 \\ &+ \sum_{n=1}^{2} \mathrm{tr}[\Phi^{(n)}\mathbf{V}^{(n)^T}] + \sum_{n=1}^{2} \mathrm{tr}[\Psi^{(n)}\mathbf{W}^{(n)^T}]. \quad (30)\end{aligned}$$

To derive multiplicative updates, we set partial derivatives of the Lagrangian function with respect to $\{\mathbf{V}^{(n)}\}_{n=1}^2$ and $\{\mathbf{W}^{(n)}\}_{n=1}^2$ equal to zero. For $n = 1, 2$, the partial derivatives are given by,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_{k,j}^{(n)}} &= [-2\mathbf{V}^{(n)^T}\mathbf{X}^{(n)}\mathbf{D}^{(n)} + 2\mathbf{V}^{(n)^T}\mathbf{V}^{(n)}\mathbf{W}^{(n)}\mathbf{D}^{(n)^T} \\ &+ 2\lambda\mathbf{W}^{(n)}\mathbf{L}^{(n)^+} - 2\lambda\mathbf{W}^{(n)}\mathbf{L}^{(n)^-} + 2\mathbf{W}^{(n)}\mathbf{D}^{(n)} \\ &- 2\mathbf{C}^{(n)}]_{k,j} + \psi_{k,j}^{(n)} = 0, \quad (31) \\ \frac{\partial \mathcal{L}}{\partial v_{i,k}^{(n)}} &= [-2\mathbf{X}^{(n)}\mathbf{D}^{(n)}\mathbf{W}^{(n)^T} + 2\mathbf{V}^{(n)}\mathbf{W}^{(n)}\mathbf{D}^{(n)}\mathbf{W}^{(n)^T}]_{k,j} \\ &+ \phi_{i,k}^{(n)} = 0, \quad (32)\end{aligned}$$

where for notation convenience we set $\mathbf{D}^{(n)} = \Delta^{(n)}\Delta^{(n)^T}$ and $\mathbf{C}^{(1)} = \mathbf{W}^{(2)}\Delta^{(1)}\Delta^{(2)^T}$ and $\mathbf{C}^{(2)} = \mathbf{W}^{(1)}\Delta^{(2)}\Delta^{(1)^T}$. As in the case of SFNMF by employing the Karush-Kuhn-Tucker conditions $\psi_{k,j}^{(n)} w_{k,j}^{(n)} = 0$ and $\phi_{i,k}^{(n)} v_{i,k}^{(n)} = 0$ the following multiplicative updates are derived:

$$\mathbf{W}_{t+1}^{(n)} = \mathbf{W}_t^{(n)} \circ \frac{\mathbf{V}_t^{(n)^T}\mathbf{X}^{(n)}\mathbf{D}_t^{(n)} + \lambda\mathbf{W}_t^{(n)}\mathbf{L}^{(n)^-} + \mathbf{C}_t^{(n)}}{\mathbf{V}_t^{(n)^T}\mathbf{V}_t^{(n)}\mathbf{W}_t^{(n)}\mathbf{D}_t^{(n)} + \lambda\mathbf{W}_t^{(n)}\mathbf{L}^{(n)^+} + \mathbf{W}_t^{(n)}\mathbf{D}_t^{(n)}}. \quad (33)$$

$$\mathbf{V}_{t+1}^{(n)} = \mathbf{V}_t^{(n)} \circ \frac{\mathbf{X}^{(n)}\mathbf{D}_t^{(n)}\mathbf{W}_{t+1}^{(n)^T}}{\mathbf{V}_t^{(n)}\mathbf{W}_{t+1}^{(n)}\mathbf{D}_t^{(n)}\mathbf{W}_{t+1}^{(n)^T}} \quad (34)$$

Similarly to the case of one sequence these updates are augmented with small positive numbers $\delta^{(n)}$ to ensure that there will not be division with zero as follows

$$\mathbf{W}_{t,r}^{(n)} = \mathbf{W}_t^{(n)} - \frac{\overline{\mathbf{W}}_t^{(n)} \circ \nabla_{W^{(n)}} F(\mathbf{V}_t^{(n)}, \mathbf{W}_t^{(n)})}{2[\mathbf{V}_t^{(n)^T}\mathbf{V}_t^{(n)}\overline{\mathbf{W}}_t^{(n)}\mathbf{D}_t^{(n)} + \lambda\overline{\mathbf{W}}_t^{(n)}\mathbf{L}^{(n)^+} + \overline{\mathbf{W}}_t^{(n)}\mathbf{D}_t^{(n)}]_{k,j} + \delta^{(n)}} \quad (35)$$

$$\mathbf{V}_{t,r}^{(n)} = \mathbf{V}_t^{(n)} - \frac{\overline{\mathbf{V}}_t^{(n)} \circ \nabla_{V^{(n)}} F(\mathbf{V}_t^{(n)}, \mathbf{W}_{t,r}^{(n)})}{2[\overline{\mathbf{V}}_t^{(n)}\mathbf{W}_{t,r}^{(n)}\mathbf{D}_t^{(n)}\mathbf{W}_{t,r}^{(n)^T}]_{i,k} + \delta^{(n)}}, \quad (36)$$

where $\mathbf{W}_{t,r}^{(n)}$ and $\mathbf{V}_{t,r}^{(n)}$ are the intermediate matrices before the normalization and the matrices $\overline{\mathbf{V}}_t^{(n)}$ and $\overline{\mathbf{W}}_t^{(n)}$ are defined as in (25) and (26), respectively. The warping matrices $\Delta^{(1)}$ and $\Delta^{(2)}$ are iteratively updated via the DTW. The iterative procedure terminates when the convergence criterion is satisfied. We used the difference of the objective function between two successive iterations as stopping criterion. The proposed algorithm for the SFNMF-TW is summarized in Algorithm 2 and its convergence can be also proved following [30].

Algorithm 2 SFNMF-TW

Data: $\mathbf{X}^{(1)} \in \mathbb{R}_+^{F_1 \times T_1}, \mathbf{X}^{(2)} \in \mathbb{R}_+^{F_2 \times T_2}, \mathbf{L}^{(1)} \in \mathbb{R}^{T_1 \times T_1}, \mathbf{L}^{(2)} \in \mathbb{R}^{T_2 \times T_2}, 1 \leq K \leq \min\{F_1, F_2, T_1, T_2\}$

Result: $\mathbf{\Delta}^{(1)} \in \{0,1\}^{T_1 \times T}, \mathbf{\Delta}^{(2)} \in \{0,1\}^{T_2 \times T}, \mathbf{V}^{(1)} \in \mathbb{R}_+^{F_1 \times K}, \mathbf{V}^{(2)} \in \mathbb{R}_+^{F_2 \times K}, \mathbf{W}^{(1)} \in \mathbb{R}_+^{K \times T_1}, \mathbf{W}^{(2)} \in \mathbb{R}_+^{K \times T_2}$

Give $\delta^{(n)} > 0$ and $\sigma^{(n)} > 0$. Initialize:
$\mathbf{V}_1^{(1)} \geq \mathbf{0}, \mathbf{V}_1^{(2)} \geq \mathbf{0}, \mathbf{W}_1^{(1)} \geq \mathbf{0}, \mathbf{W}_1^{(2)} \geq \mathbf{0}, t = 1$
while *not converged* **do**
  $(\mathbf{\Delta}_t^{(1)}, \mathbf{\Delta}_t^{(2)}) \leftarrow \mathrm{DTW}(\mathbf{W}_t^{(1)}, \mathbf{W}_t^{(2)})$
  for *n=1 to 2* **do**
    Update $\mathbf{W}_{t,r}^{(n)}$ according to Eq. (35)
    Update $\mathbf{V}_{t,r}^{(n)}$ according to Eq. (36)
    Normalise $\mathbf{W}_{t,r}^{(n)}$ and $\mathbf{V}_{t,r}^{(n)}$ to $\mathbf{V}_t^{(n)}$ and $\mathbf{W}_t^{(n)}$ respectively, so that the sum of $\mathbf{V}_t^{(n)}$ columns is one. In case where the whole column is zero, then this and the corresponding row in $\mathbf{V}^{(n)}$ do not change.
  $t \leftarrow t + 1$

Empirical convergence has been always observed in all tested videos both in terms of the cost function (30), as well as for the DTW step in (6). Fig. 3 shows the averaged convergence curves of SFNMF-TW versus the number of iterations in both MMI and UNS datasets. Furthermore, Fig. 4 shows the evolution of the DTW error term of the cost function with respect to iterations.

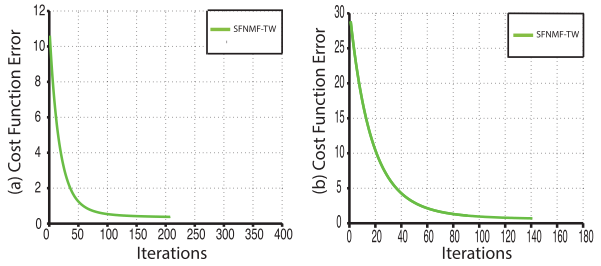

Fig. 3. Convergence curve on (a) MMI database (b) UNS database.

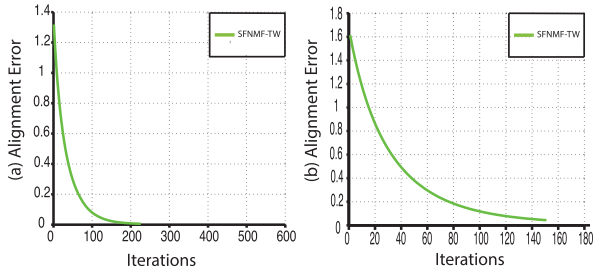

Fig. 4. DTW error term in (a) MMI database (b) UNS database.

V. Experimental Results

The performance of the proposed methods is assessed by conducting experiments on the MMI [25], [26] and the UvA-Nemo Smile (UNS) [27] databases. The MMI [25], [26] consists of videos with posed FAUs while the UNS contains videos with posed and spontaneous smiles.

The MMI contains more than 400 videos annotated in terms of FAUs and the temporal segments in which a subject performs one or more FAUs in terms of neutral-onset-apex-offset-neutral indicators. We used 351 of those videos and we tracked 68 facial landmarks using a variant of the Supervised Descent Method (SDM) [31]. The tracked landmarks were used in order to align and scale the frames to a fixed size template of 169×171 pixels. The relevant FAUs used for each region of the face are as follows:

- **Mouth**: Upper Lip Raiser, Nasolabial Deepener, Lip Corner Puller, Cheek Puffer, Dimpler, Lip Corner Depressor, Lower Lip Depressor, Chin Raiser, Lip Puckerer, Lip stretcher, Lip Funneler, Lip Tightener, Lip Pressor, Lips part, Jaw Drop, Mouth Stretch and Lip Suck
- **Eyes**: Upper Lid Raiser, Cheek Raiser, Lid Tightener, Nose Wrinkler, Eyes Closed, Blink, Wink, Eyes turn left and Eyes turn right
- **Brows**: Inner Brow Raiser, Outer Brow Raiser and Brow Lowerer.

UvA-Nemo Smile database contains more than 1000 smile videos (597 spontaneous and 643 posed) from 400 subjects. The database does not provide annotations with regards to temporal segments. Hence, we annotated 100 videos in total, 50 displaying posed and 50 displaying spontaneous smiles, in terms of temporal segments. Furthermore, we used the same algorithm to track 68 facial landmarks and align the facial images.

A. Unsupervised Analysis of Facial Temporal Dynamics in One Sequence

In this section, the performance of the SFNMF is compared against that of the NMF, the GNMF [32], and the SFA for unsupervised facial behaviour analysis. More precisely, we investigated how effectively each method can detect the transitions between the temporal phases (i.e., Neutral-Onset-Apex-Offset) during different facial AUs activation.

The parameters of each method were tuned by using a validation set. For GNMF we considered a 5-nearest neighbors graph to capture the local geometric structure of data, a $0-1$ weighting system for defining the weight matrix and set parameter λ that regulates the contribution of the two parts in GNMF cost function to 150. Finally, for all algorithms we considered projection to a subspace of equal dimensionality which was set to 50 and 250 and the step sizes of the modified updates rules were set $\delta = \sigma = 10^{-8}$.

To facilitate the comparison between the results of each method and the ground truth, we map the recovered latent space by each method to the temporal phases of AUs. This is done by finding for each method the slowest varying latent feature. To do so, we compute the first order derivative for each obtained latent variable and select the one that minimizes: $\text{argmin}_i \, \mathbf{w}_i \mathbf{L} \mathbf{w}_i^T$. We should note that since SFA introduces an ordering to the derived latent variables sorted by their temporal slowness, we simply acquire the first identified latent feature which corresponds to the slowest varying one.

Fig. 5 shows the performance of the examined methods in terms of capturing the AU temporal phases and in terms of extracting accurate part based representations on two video sequences displaying the activation of two different AUs. More precisely, the results presented in Fig. 5(a) correspond to a video sequence where the subject performs AU 26 (i.e. Jaw Drop), while results shown in Fig. 5(b) correspond to the activation of AU 43 (i.e. eyes closed). In each plot the ground truth (green curve) instances when the AUs temporal phases transition appear are highlighted with red marks. As can be observed in both videos the proposed method outperforms both GNMF and SFA since it detects the temporal phases more accurately while NMF was not able to detect the transition between the AU's temporal phases on both videos. Moreover, Fig. 5(a) shows the basis images ($\mathbf{V}$) corresponding to the features that best capture the dynamics of the AU 26. As can be seen the extracted basis from SFNMF depict better the activated facial part (mouth) related to AU 26 compared to other NMF-based algorithms and SFA. Finally, in Fig. 5(b) we can observe the corresponding basis for the eye-related AU (AU 43) where it is obvious that the part-based decomposition from SFNMF produced the better basis image.

Even though the GNMF does not explicitly capture the temporal dynamics in the visual sequence, the nearest neighbours of each datum (which are encoded in the k-NN graph in the GNMF) are usually successive video frames. Therefore, the temporal information is encoded implicitly. This fact justifies the good performance of the GNMF. Furthermore, regarding SFA, it is clear from Fig. 5(b) that the SFA's basis image differs significantly from the one obtained by the

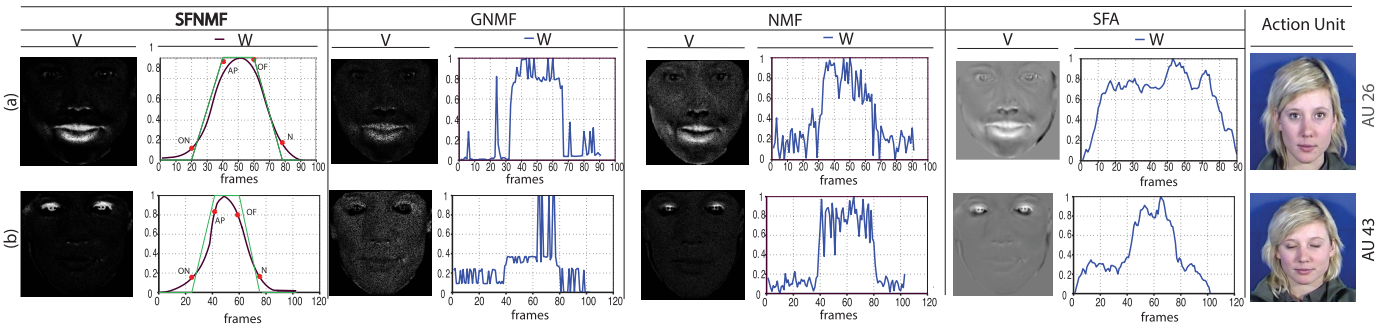

Fig. 5. Extracted features (**W**) along with the corresponding basis (**V**) by applying the SFNMF, NMF, GNMF and SFA on a video sequence from the MMI database on a subject performing: (a) Jaw Drop (AU 26) (b) Eyes Closed (AU 43). The red marks indicate the annotated ground truth (green curves) where the AU temporal phase changes.

TABLE I

ERROR BETWEEN THE EXTRACTED FEATURES AND GROUND TRUTH ANNOTATIONS FOR EACH TEMPORAL PHASE ON THE MMI DATABASE FOR $K = 50$. THE RESULTS COMPARE THE PERFORMANCE OF THE SFNMF AGAINST GNMF, NMF AND SFA ON GROUND TRUTH SHAPE

	Neutral			Onset			Apex			Offset			Overall			Total
Method	Mouth	Eyes	Brows	Mouth	Eyes	Brows	Mouth	Eyes	Brows	Mouth	Eyes	Brows	Mouth	Eyes	Brows	
SFNMF	2.856	11.492	3.978	**0.305**	**0.414**	0.322	**1.714**	**1.320**	2.111	**0.406**	**0.686**	0.521	**0.290**	**0.613**	**0.397**	**0.379**
GNMF	2.556	10.595	3.267	0.341	0.442	0.369	2.639	2.003	**2.052**	0.417	0.702	**0.383**	0.703	1.037	0.711	0.804
NMF	**0.803**	**0.926**	**0.907**	2.979	1.997	2.529	26.144	15.824	19.725	2.367	2.078	1.855	1.072	0.760	0.744	0.960
SFA	10.186	17.669	8.535	0.725	0.446	0.523	1.286	1.155	1.582	1.046	1.206	0.957	0.404	0.921	0.534	0.544
SFA (points)	3.958	12.556	4.353	0.424	0.451	**0.310**	3.475	2.463	1.974	0.601	0.853	0.515	1.165	1.736	1.051	1.284

TABLE II

ERROR BETWEEN THE EXTRACTED FEATURES AND GROUND TRUTH ANNOTATIONS FOR EACH TEMPORAL PHASE ON THE MMI DATABASE FOR $K = 250$. THE RESULTS COMPARE THE PERFORMANCE OF THE SFNMF AGAINST GNMF AND NMF ON GROUND TRUTH SHAPE

	Neutral			Onset			Apex			Offset			Overall			Total
Method	Mouth	Eyes	Brows	Mouth	Eyes	Brows	Mouth	Eyes	Brows	Mouth	Eyes	Brows	Mouth	Eyes	Brows	
SFNMF	2.6735	10.8797	3.4993	0.3486	**0.3870**	**0.3940**	1.8441	**1.1461**	**1.6430**	**0.3701**	**0.7307**	0.3715	**0.2552**	**0.5116**	**0.2691**	**0.2971**
GNMF	**2.2452**	11.1410	3.0440	**0.2870**	0.4461	0.3551	**1.6072**	1.3495	2.4155	0.3791	0.8917	**0.3208**	0.7705	0.9402	0.8698	0.822
NMF	3.0014	**6.6520**	**2.2651**	0.7929	0.5198	0.5880	8.7436	3.8599	5.1086	0.7448	0.8333	0.5363	1.7884	1.9761	1.3906	1.7852

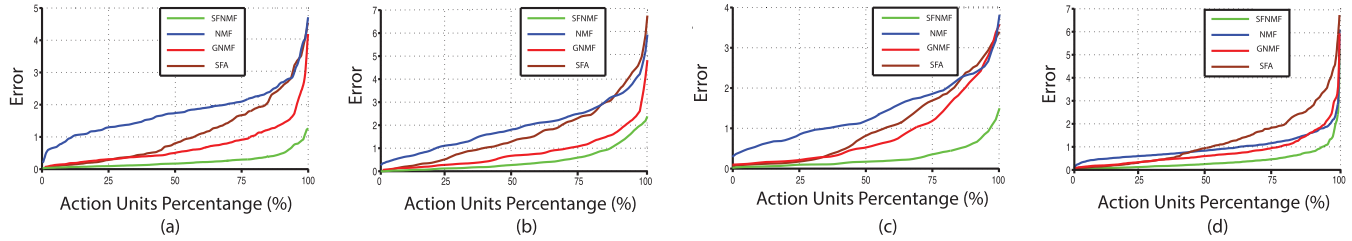

Fig. 6. Overall error between the extracted features and the annotated ground truth on the MMI database. The plots compare the performance of the SFNMF against SFA, NMF and GNMF. (a) Mouth-related AUs (b) Eyes-related AUs (c) Brows-related AUs (d) All AUs.

NMF-based techniques. This is because, SFA does not enforces nonegativity constraints and thus produces holistic representation which makes latent features difficult to be interpreted. On the other hand, the latent features of NMF are much more noisy compared to that obtained by the GNMF and the SFNMF due to the fact that NMF lacks of smoothing constraint

Table I and II summarize the results for each temporal phase where we provide results for Mouth-related AUs, Eyes-related AUs and Brows-related AUs separately for the MMI database. Specifically, it reports the mean error for each temporal phase along with the overall error for the whole performed AU and the total error for all of the AUs. For measuring the error, we applied the DTW algorithm between the extracted features and the ground truth. The presented results indicate that the SFNMF algorithm performs better than the other methods on the unsupervised detection of the temporal phases of FAUs, almost in all temporal phases and for all relevant regions of the face for both $K = 50$ and $K = 250$ as can be seen in Table I and Table II respectively. Additionally, comparing these tables we can notice that there was not any significant improvement in the performance of the applied methods when setting the number of the desired features to be extracted to 250, as someone should expect. This is attributed to the fact that extracting 50 features was sufficient to preserve more than 90% of the original sequences' energy. The overall performance of the examined methods is better visualized in Fig. 6 which reports the error versus the percentage of the videos for each region of the face separately. For instance, Fig. 6(a) shows the error for all of the AUs performed by the

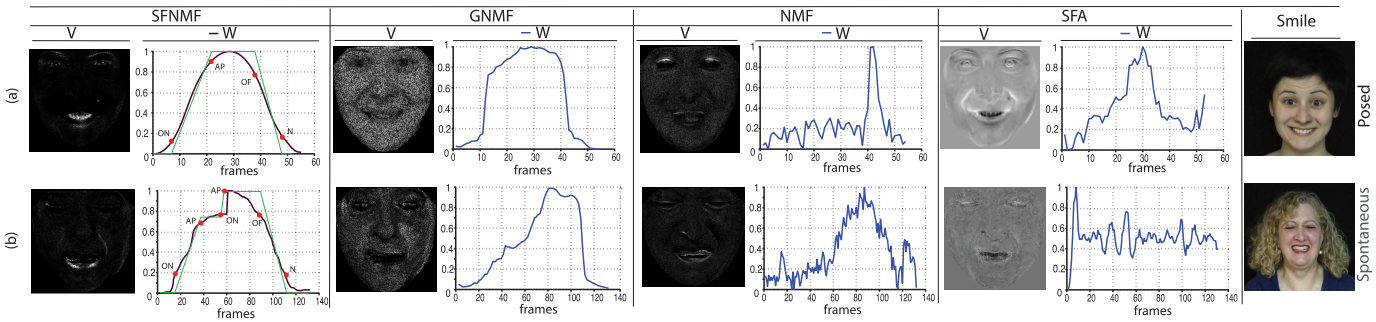

Fig. 7. Extracted features (**W**) along with the corresponding basis (**V**) by applying the SFNMF, NMF, GNMF and SFA on a video sequence from the UNS database on a subject performing: (a) Spontaneous Smile (b) Posed Smile. The red marks indicate the annotated ground truth where the AU temporal phase changes.

TABLE III

ERROR BETWEEN THE EXTRACTED FEATURES AND GROUND TRUTH ANNOTATIONS FOR EACH TEMPORAL PHASE ON THE UNS DATABASE FOR $K = 50$. THE RESULTS COMPARE THE PERFORMANCE OF THE SFNMF AGAINST GNMF AND NMF ON GROUND TRUTH SHAPE

	Neutral		Onset		Apex		Offset		Overall	
Method	Posed	Spontaneous	Posed	Spontaneous	Posed	Spontaneous	Posed	Spontaneous	Posed	Spontaneous
SFNMF	**0.1127**	0.0220	**0.3271**	**1.1798**	**2.9787**	**8.9824**	**0.3225**	**0.6845**	**0.0658**	**0.0575**
GNMF	0.1284	**0.0098**	0.6199	1.5147	4.7353	14.5067	0.4869	0.8903	0.0834	0.0767
NMF	0.4590	0.1445	4.2596	5.0809	36.6849	73.3768	3.1687	3.3437	1.4057	2.0399
SFA	0.8086	0.2285	0.9803	2.0762	7.9692	14.1013	0.9194	1.4587	0.5407	0.3850
SFA (points)	0.9691	0.7117	0.4267	1.1641	6.1399	14.0671	0.4982	1.4966	0.9080	1.3329

mouth, Fig. 6(b) shows the error from the Eyes-related AU, Fig. 6(c) from the Brows-related AU and Fig. 6(d) shows the overall error for all the AUs.

Finally, Table IV reports the average correlation accuracy of the bases obtained by the applied methods for the MMI database. Specifically, the reported results were obtained by measuring the correlation of the activated facial parts between the produced basis image and the original one. The results verify that in average the bases extracted by the SFNMF algorithm were capable of capturing more accurately the relevant activated facial parts to the performed AU. Additionally, by inspecting the Table IV, we observe that the results are consistent with that of in Table I and II, indicating that the features (**W**) are faithful representatives for evaluating qualitatively the extracted bases.

Next we test the performance of the examined methods in UNS database. Specifically, we applied the methods on 50 spontaneous and 50 posed smile videos. The performance was measured by applying DTW between the extracted features and the annotated ground truth.

Fig. 7 compares the extracted features, of the examined methods, that best capture the temporal phases when a subject performs a posed Fig. 7(a) and a spontaneous smile Fig. 7(b), from the UNS database, respectively. In addition, next to each feature we can see the corresponding basis image. The features (**W**) in Fig. 7(a) indicate that the SFNMF algorithm outperform the other methods since it detects the dynamics of the smile more accurately and captures the temporal phases more smoothly. The same occurs and in the case where the subject performs spontaneous smile (Fig. 7(b)) which is more challenging to capture its dynamics accurately due to the fact that it is consisting of multiple temporal phases (two apex and two offset phases). Moreover, the basis images extracted from SFNMF were the only ones that depict the correct activated facial part (i.e. mouth) for both Spontaneous and Posed smiles, as can be seen in Fig. 7 (a) and Fig. 7 (b), respectively.

TABLE IV

QUANTITATIVE RESULTS AMONG THE BASES OF THE SFNMF, GNMF, NMF, AND SFA ON THE MMI DATABASE. THE RESULTS COMPARE THE AVERAGE CORRELATION BETWEEN THE EXTRACTED BASES AND THE ORIGINAL IMAGES SEPARATELY FOR THE MOUTH, EYES AND BROWS RELATED AUS

Related AUs	Method	Correlation
Mouth	**SFNMF**	**0.3762**
	GNMF	0.3320
	NMF	0.2519
	SFA	0.3604
Eyes	**SFNMF**	**0.4476**
	GNMF	0.3657
	NMF	0.3077
	SFA	0.2776
Brows	**SFNMF**	**0.6463**
	GNMF	0.4965
	NMF	0.3977
	SFA	0.1375

The overall results are summarized in Table III which reports the mean error for each temporal phase and for both Posed and Spontaneous instances. Similarly to the MMI experiments, the results indicate that the SFNMF outperforms the other methods on the unsupervised detection in almost all temporal phases. The performance of all videos for each of the compared methods can be better seen in Fig. 8. Particularly, Fig. 8(a) shows the error for all of the posed and Fig. 8(b)

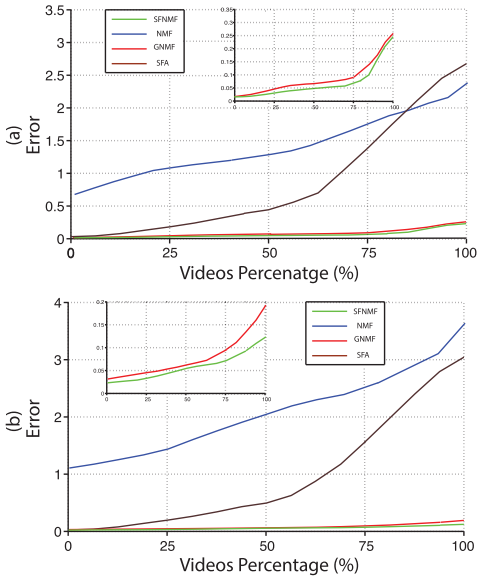

Fig. 8. Overall error between the extracted features and the annotated ground truth on the UNS database. The plots compare the performance of the SFNMF against SFA, NMF and GNMF. (a) Posed smiles (b) Spontaneous smiles.

TABLE V

QUANTITATIVE RESULTS AMONG THE BASES OF THE SFNMF, GNMF, NMF, AND SFA ON THE UNS DATABASE. THE RESULTS COMPARE THE AVERAGE CORRELATION BETWEEN THE EXTRACTED BASES AND THE ORIGINAL IMAGES SEPARATELY FOR THE SPONTANEOUS AND POSED SMILES

Smiles	Method	Correlation
Spontaneous	**SFNMF**	**0.2950**
	GNMF	0.2578
	NMF	0.2014
	SFA	0.2815
Posed	**SFNMF**	**0.3280**
	GNMF	0.3061
	NMF	0.1783
	SFA	0.3011

shows the error for all of the spontaneous smiles, respectively. For better clarity, we zoomed in the performance of the SFNMF and GNMF as can be seen in the top right corner for both graphs. Finally, the relative results for the quantitatively evaluation of the bases for the UNS database are presented in Table V.

B. Temporal Alignment of Facial Events for Two Sequences

In this section we provide experimental results on aligning pairs of videos from the MMI and UNS databases, where the same AU is activated. The aim of the experiment is two fold: (a) to test various dimensionality reduction and component analysis methods for temporal alignment and (b) to compare the performance of simultaneous decomposition and alignment procedures with the ones that first decompose the signals and then apply DTW. To this end, we compare a number of component analysis that are relevant to the problem, namely **(a)** CTW and PCA+CTW. **(b)** Deterministic Slow Feature analysis (SFA) [6] plus DTW (SFA+DTW). **(c)** NMF plus DTW (NMF+DTW). **(d)** Joint NMF and DTW, which produced by the optimisation of the proposed problem (30) setting $\mathbf{L}^{(1)}$ and $\mathbf{L}^{(2)}$ to zero. **(e)** Graph Regularised NMF (GNMF) [32] plus DTW. **(f)** Joint GNMF and DTW, which produced by the optimisation problem (30) setting $\mathbf{L}^{(1)}$ and $\mathbf{L}^{(2)}$ equal to a graph Laplacian. **(g)** Graph regularised NMF using the graph in (10) plus DTW (so-called SFNMF [33]). **(h)** The proposed SFNMF-TW.

We have also tested the probabilistic models in [4] and [11] but due to the very high dimensionality of the inputs their performance was not satisfactory.[1] Finally, we tested the methodology presented in [17] but since our videos did not contain any gross errors we did not observe any improvement over CTW. Thus, we do not report the performance for the methods [4], [11], [17] in order not to clutter our graphs. To the best of our knowledge such thorough comparison of component analysis techniques applied for temporal alignment has not be conducted before in the literature.

We used a two pairs of videos (one for MMI and one for UNS), not used in the test phase, as a validation step to fine tune the involved parameters. In the case of GNMF methods we have tested various approaches to build the graph Laplacian. That is, we used the heat kernel, dot-product kernels, 0–1 weighting etc. and we tested various neighbourhood sizes. The best was a 5-nearest neighbours graph to capture the local geometric structure of data using a $0-1$ weighting system for defining the weights. For all joint NMF and DTW techniques the parameter λ that regulates the contribution of the two parts cost function was set equal to 100 and the step sizes δ and σ equal to 10^{-8}. Furthermore, the dimensionality of the latent space was set equal to 50. The dimensionality of CTW and SFA was set to $K = 15$ and $K = 30$ for MMI and UNS, respectively. It is worth noting that for CTW we had always to apply a PCA step beforehand otherwise the algorithm performed poorly. All joint algorithms were allowed to run until convergence which was determined by monitoring the objective function improvement between successive iterations. The experiments were conducted in 485 pairs of videos from MMI depicting the same FAUs and 100 pairs of posed and 100 pairs of spontaneous smiles from UvA-Nemo.

1) Experiments in MMI: We present the experiments in MMI according to the region of the face the FAU depicts (i.e., mouth, eyes and brows related-AUs separately), as well as overall results from all FAUs. In the first set of experiment we measured the objective alignment error (i.e., the error produced by the DTW normalised with the number of dimensions) for each of the compared method with respect to the percentage of aligned pairs of videos.[2] The evolution of the errors is plotted in Fig. 11.

As can be observed all the joint decomposition and alignment procedures (solid lines) vastly outperformed the application of alignment algorithms on the features produced by the signal decomposition (for instance please inspect the difference in the performance by comparing the results

[1] Actually in [4] and [11] the aligned signals were of very low-dimensionality.

[2] For example the point (60%,1) of graph mean that 60% of the pairs have error lower than or equal to 1

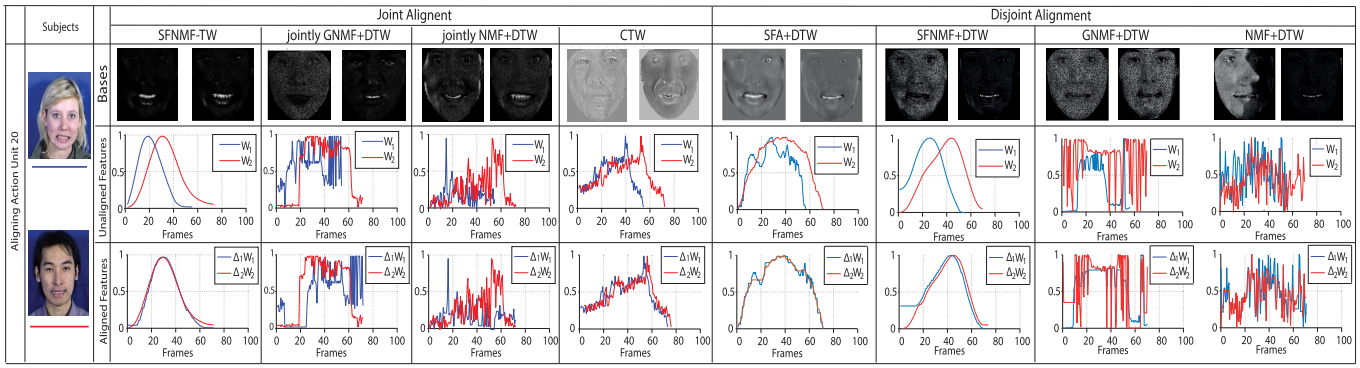

Fig. 9. Aligning the AU20 (Lip stretcher) performed by two different subjects. (Top row) Basis obtained by the tested methods (Mid row) Original extracted features (Bottom row) Aligned extracted features.

TABLE VI
AVERAGE ALIGNMENT ERROR IN MMI DATABASE

Alignment Error				
Method	Mouth	Eyes	Brows	Overall
SFNMF-TW	**0.0312**	**0.0428**	**0.0365**	**0.0389**
Jointly GNMF+DTW	0.1784	0.1912	0.1586	0.1873
Jointly NMF+DTW	0.8939	1.2023	1.0146	1.1096
SFNMF+DTW	0.2039	0.2492	0.1607	0.1962
GNMF+DTW	0.2119	0.2897	0.1886	0.2127
NMF+DTW	1.3686	1.7225	1.459	1.6154
SFA+DTW	0.9171	1.196	1.1364	1.1116
CTW	0.7608	0.9758	0.7175	0.8016

between the NMF+DTW (blue dashed line) and jointly NMF+DTW (blue solid line) in Fig. 11(a)). Finally, The average DTW error for all the videos of MMI database is summarised in Table VI. As can be seen the proposed methodology vastly outperforms all other tested methods.

The DTW error provides as only an indication of the efficacy of the tested algorithms. Hence, we further evaluated the accuracy of each algorithm by using a robust metric used in recent works [4]. In more detail, lets assume two videos, with features $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ and AU annotations $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$. All tested methods using these features recover the alignment matrices $\Delta^{(1)}$ and $\Delta^{(2)}$. By applying these matrices on the AU annotations (i.e., $\mathbf{A}^{(1)}\Delta^{(1)}$ and $\mathbf{A}^{(2)}\Delta^{(2)}$) we know to which temporal phase of the AU each aligned frame of each video corresponds to. Therefore, for a given temporal phase (e.g., neutral), we have a set of frame indices which are assigned to the specific temporal phase in video 1, $\mathcal{N}^{(1)p}$ and video 2, $\mathcal{N}^{(2)p}$. The accuracy is then estimated as

$$\frac{|\mathcal{N}^{(1)p} \cap \mathcal{N}^{(2)p}|}{|\mathcal{N}^{(1)p} \cup \mathcal{N}^{(2)p}|} \tag{37}$$

which essentially corresponds to the ratio of correctly aligned frames to the total duration of the temporal phase p across the aligned videos.

The alignment accuracy (37) obtained for all the examined methods and for all temporal segments in the MMI database is shown in Fig. 12. As before the results are presented for various facial parts (mouth, eyes and brows) separately. The darker colours are use to show the accuracy of joint decomposition and alignment techniques, while the light colours are used for techniques that treat alignment and decomposition as separate steps (e.g., light red represents GNMF+DTW while darker red joint GNMF and DTW). It can be verified that the proposed methodology outperforms all others.

Finally for better inspection of the bases and the latent features we provide two experiments of aligning pairs of videos where the subjects perform the same AU. Specifically, Fig. 9 compares the obtained results, when aligning two videos of two subjects performing FAU 20 (Lip stretcher), employing the proposed framework for joint alignment with the obtained results when aligning the sequences after having extracted their features (disjoint alignment). The top row shows bases images of the part-based decomposition extracted by the tested methods. As can be seen the bases extracted by the proposed methodology (SFNMF-TW) are the only ones who both correspond to the common facial part that is activated (i.e, mouth) compared to the other joint methods. Additionally, by comparing the extracted bases of the joint methods with the disjoint ones we notice that the bases of the disjoint methods are unrelated to each other. This do not occur in the methods that perform joint decomposition and alignment, regardless their quality, since those two procedures are applied simultaneously and are dependent. The mid and the bottom rows plot the two latent features, one for each behavioural sequence (blue for the left and red for the right), over the whole video for both joint and disjoint methods before and after applying the alignment process, respectively. Comparing the joint methods to each other can be verified that the proposed methodology (SFNMF-TW) achieves better alignment since it provides smoother latent spaces. Finally, by comparing the latent features between the joint and the disjoint methods we can see that all the joint methods managed to extract better latent features (less noisy). This is attributed to the iterative nature of this framework which allows both the temporal segmentation and the alignment process to be gradually improved.

The second indicative example can be seen in Fig. 10 where the subjects perform AU 5 (Upper Lid Raiser). As can be seen from the first row, the features extracted from the proposed methodology detect the transitions between the temporal phases (Onset, Apex, Offset and Neutral) during the

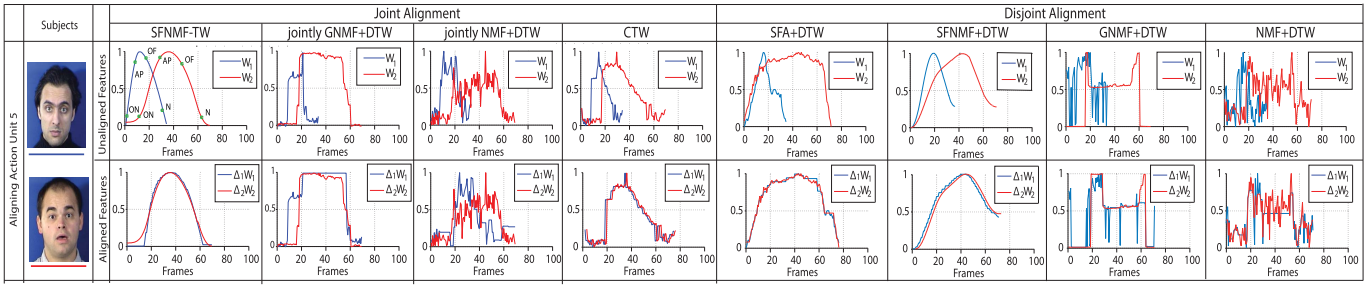

Fig. 10. Aligning the AU5 (Upper Lid Raiser) performed by two different subjects.(first row) Original extracted features (Second row) Aligned extracted features. The green marks indicate the annotated ground truth where the AU temporal phase changes (N - Neutral phase, ON - Onset phase, AP - Apex phase, OF - Offset phase).

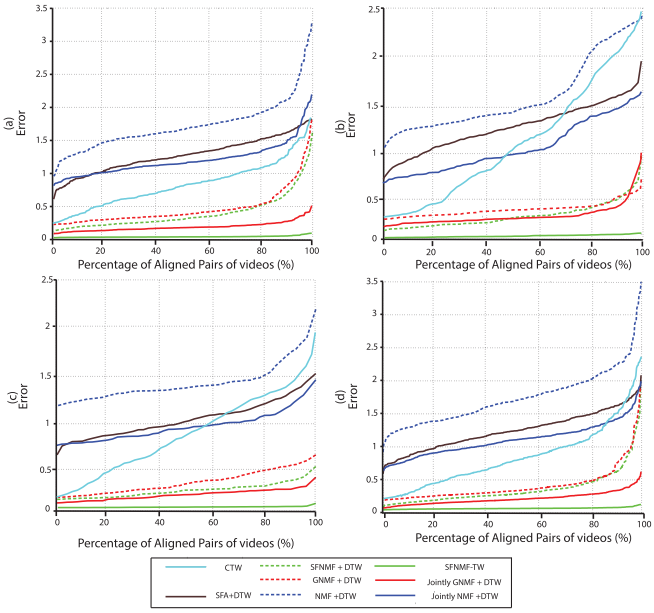

Fig. 11. Overall results obtained when aligning pairs of videos from MMI database where the same AU is activated in (a) Mouth-related AUs (b) Eyes-related AUs (c) Brows-related AUs (d) All the AUs.

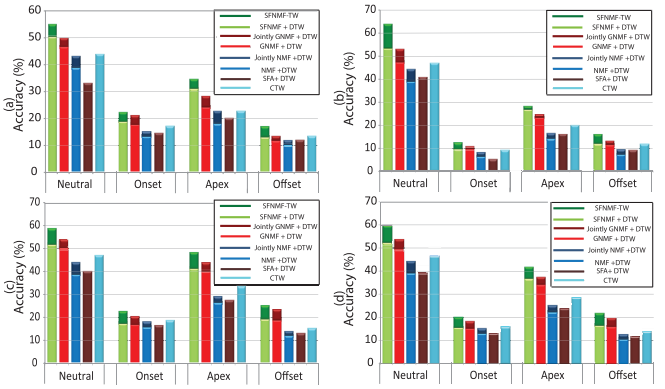

Fig. 12. Accuracy of the tested methods in alignment tasks for all temporal phases in (a) Mouth-related AUs (b) Eyes-related AUs (c) Brows-related AUs (d) all AUs.

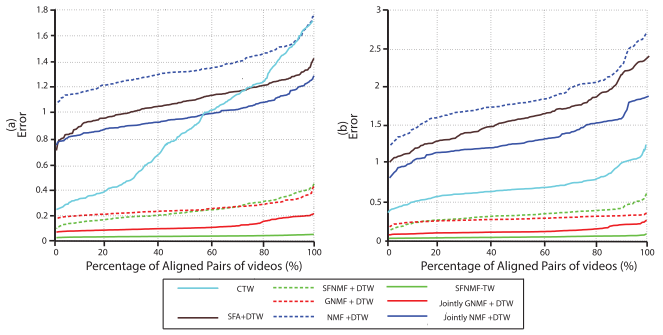

Fig. 13. DTW alignment error for the pairs of videos of the UNS database for (a) posed smiles (b) spontaneous smiles.

AU activation more accurately and more smoothly compared to other joint methods and hence, these features can be better be aligned as can been observed in the second row of the Fig. 10. Moreover, similar to the previous qualitative experiment, the methods which are subject to independent alignment produce more noisy feature components compared to the respective joint methods (e.g inspect the features of jointly GNMF+DTW vs GNMF+DTW).

2) Experiments in UNS Database: A similar experimental setup was used in UNS database to test all the alignment algorithms in videos displaying more complex expressions, that is of both posed and spontaneous smiles. Fig. 13 plots the DTW error curves versus the percentage of the videos for both posed and spontaneous smiles, while Table VII provides the average DTW errors.

Additionally, Fig. 14 provides the corresponding graphs of the alignment accuracy metric (37). By comparing these figures we notice that the results in alignment error for the spontaneous facial displays (Fig. 13 (b)) are not in line with the results which measure the alignment error (Fig. 14 (b)) compared to the respective results for the posed facial displays. Specifically, by inspecting the Fig. 13 (b) we can see that GNMF outperforms SFNMF when combining with DTW in terms of the alignment error. This is mainly due to the fact that the spontaneous facial instances consist of more that one temporal phases which makes the alignment error in such cases not the most accurate mean to evaluate the performance of the alignment methods. Therefore, the most precise way to measure the performance of the examined methods is by measuring their accuracy when aligning each temporal phase

TABLE VII
AVERAGE ALIGNMENT IN UNS DATABASE

Alignment Error		
Method	Spontaneous	Posed
SFNMF-TW	**0.0619**	**0.0424**
Jointly GNMF+DTW	0.1325	0.1186
Jointly NMF+DTW	1.4385	0.9736
SFNMF+DTW	0.1888	0.1892
GNMF+DTW	0.1636	0.2073
NMF+DTW	1.9014	1.3343
SFA+DTW	1.6868	1.0843
CTW	0.7117	0.8719

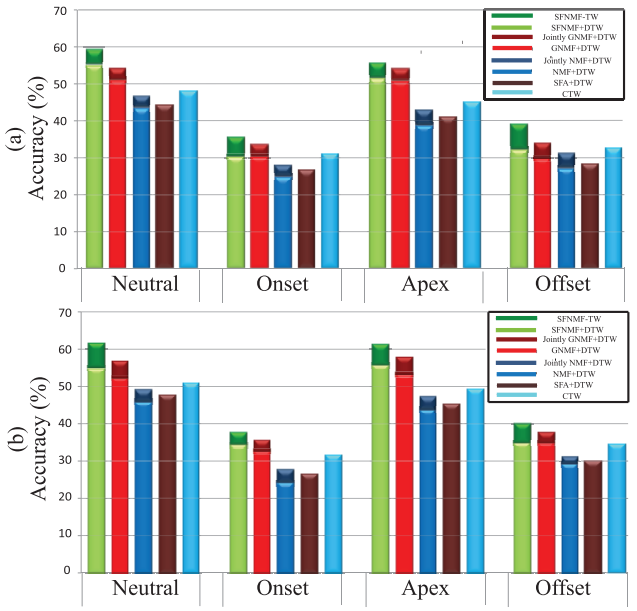

Fig. 14. Overall accuracy obtained when comparing the examined methods in alignment tasks for all temporal phases in (a) Posed Smiles (b) Spontaneous Smiles.

separately. Such evaluation is depicted in Fig. 14 (b). More precisely, this figure shows clearly that every temporal phase is better aligned when combining SFNMF with DTW compared to any other disjoint method, which in turn indicates that the features produced by SFNMF describe the temporal evolution of the spontaneous facial activities more accurately.

Finally, the experiments in UNS database demonstrate once more that it is better to perform joint decomposition and alignment than applying these two tasks separately and that the proposed SFNMF-TW vastly outperforms both the tested methods and state-of-the-art methods such as CTW.

VI. CONCLUSION

In this paper, the SFNMF has been proposed in order to learn slow varying parts-based representations of time varying facial sequences. The proposed method minimizes the data reconstruction error and the temporal variance of the derived latent features. The SFNMF has been applied in unsupervised facial behaviour dynamics analysis. Furthermore, the SFNMF has been extended in order to hadle temporally misaligned video sequecnes depicting the same facial event. This approach has been tested on temporal aliment of facial behaviour. Both the SFNMF and the SFNMF-TW outperforms the methods that they have been compared to. Finally, it is worth mentioning that an exciting area for further research on the topic is how to design non-linear dynamical systems, e.g. using deep neural network architectures, which take into account the constraints we have imposed on our model for unsupervised extraction and analysis of the dynamics of facial behaviour.

APPENDIX I
CONVERGENCE ANALYSIS OF ALGORITHM 1

Here in, we discuss the convergence of the algorithm 1 beginning by showing that from $\mathbf{W}_t$ to $\mathbf{W}_{t,r}$, the components do not satisfy the KKT conditions change and the (14) is strictly decreased while elements satisfying KKT conditions do not.

Our analysis will make use of the auxiliary function similar to that introduced by Lee and Seung [5] as follows

$$G(\mathbf{w}, \mathbf{w}_t) \equiv \overline{F}(\mathbf{w}_t) + (\mathbf{w} - \mathbf{w}_t)^T \nabla \overline{F}(\mathbf{w}_t) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_t)^T \mathbf{D}(\mathbf{w} - \mathbf{w}_t) \quad (38)$$

where $\mathbf{D}$ is a diagonal matrix and in our case is given by

$$\mathbf{D}_{kk} = \frac{[\mathbf{V}^T\mathbf{V}\mathbf{w}_t + \lambda \mathbf{w}_t \mathbf{L}^+]_k}{[w_t]_k} \quad \forall \quad k = 1, \ldots, K \quad (39)$$

and the function $\overline{F}(\mathbf{w})$ is given by

$$\overline{F}(\mathbf{w}) = \frac{1}{2}(||\mathbf{x} - \mathbf{V}\mathbf{w}||^2 + \lambda \mathrm{tr}[\mathbf{w}\mathbf{L}\mathbf{w}^T]) \quad (40)$$

where $\mathbf{x}$ is a column of $\mathbf{X}$ and $\mathbf{V} = \mathbf{V}_t$ assuming that $\mathbf{V}_t$ is fixed.

The importance of the auxiliary function can be perceived due to the following lemma

Lemma 1: If G is an auxiliary function of F, then F is nonincreasing under the update

$$\mathbf{w}_t = \underset{\mathbf{w}}{\operatorname{argmin}}\, G(\mathbf{w}, \mathbf{w}_t). \quad (41)$$

Proof:

$$\overline{F}(\mathbf{w}) \leq G(\mathbf{w}, \mathbf{w}_t) \leq G(\mathbf{w}_t, \mathbf{w}_t) = \overline{F}(\mathbf{w}_t) \quad (42)$$

□

Minimising $G(\mathbf{w}, \mathbf{w}_t)$ with respect to $\mathbf{w}$ leads to the update rule in 21. In addition, if $\overline{F}(\mathbf{w}_t)_k \neq 0$, then $[w_{t,r}]_k \neq [w_t]_k$. The limitation of this auxiliary function is that it is not well defined when $[w_t]_k$. To address that and also to deal with indices not satisfying KKT conditions we define the following auxiliary function

$$\overline{G}(\mathbf{w}, \mathbf{w}_t) \equiv \overline{F}(\mathbf{w}_t) + (\mathbf{w} - \mathbf{w}_t)_I^T \nabla \overline{F}(\mathbf{w}_t)_I + \frac{1}{2}(\mathbf{w} - \mathbf{w}_t)_I^T \overline{\mathbf{D}}_{II}(\mathbf{w} - \mathbf{w}_t)_I \quad (43)$$

where

$$I \equiv \{k | [w_t]_k > 0, \nabla \overline{F}(\mathbf{w}_t)_k \neq 0 \; or \; [w_t]_k = 0, \nabla \overline{F}(\mathbf{w}_t)_k < 0\} = \{k | [\overline{w}_t]_k > 0, \nabla \overline{F}(\mathbf{w}_t)_k \neq 0\} \quad (44)$$

and $\overline{\mathbf{D}}_{II}$ is a diagonal matrix with elements

$$\overline{\mathbf{D}}_{II} \equiv \begin{cases} \frac{[\mathbf{V}^T\mathbf{V}\overline{\mathbf{w}}_t + \lambda\overline{\mathbf{w}}_t\mathbf{L}^+]_k + \delta}{[\overline{w}_t]_k}, & if\ k \in I \\ 0, & if\ k \notin I \end{cases} \tag{45}$$

This auxiliary function is well defined when $[w_t]_k = 0$ and $\nabla\overline{F}(\mathbf{w}_t)_k < 0$ and $[w_t]$ can be changed as well. Finally, the only thing left is to ensure that (43) satisfies the nonincreasing property (42) which is shown by the following theorem

Theorem 1: For given δ and σ, $\mathbf{w}_t$ be a column of $\mathbf{W}_t$ in Algorithm 1 and $I' \equiv \{1, \ldots, k\}\backslash I$, then

$$\underset{\mathbf{w}_I}{\operatorname{argmin}}\, \overline{G}(\mathbf{w}, \mathbf{w}_t) = (\mathbf{w}_t)_I - \overline{D}_{II}^{-1}\nabla\overline{F}(\mathbf{w}_t)_I \tag{46}$$

for the update rule $\mathbf{w}_{t,r}$ given by (27) it holds that

$$(\mathbf{w}_{t,r})_I = \underset{\mathbf{w}_I}{\operatorname{argmin}}\, \overline{G}(\mathbf{w}, \mathbf{w}_t)\ and\ (\mathbf{w}_{t,r})_{I'} = (\mathbf{w}_{t,r})_{I'} \tag{47}$$

and

$$\overline{F}(\mathbf{w}_{t,r}) \leq \overline{G}(\mathbf{w}_{t,r}, \mathbf{w}_t) \leq \overline{G}(\mathbf{w}_t, \mathbf{w}_t) = \overline{F}(\mathbf{w}_t) \tag{48}$$

Proof: As $\mathbf{D}_{II}$ is positive definite, $\overline{G}(\mathbf{w}, \mathbf{w}_t)$ is a strictly convex function of $\mathbf{w}_I$, and has a unique minimum satisfying

$$\overline{\mathbf{D}}_{II}(\mathbf{w} - \mathbf{w}_t)_I + \nabla\overline{F}(\mathbf{w}_t)_I = \mathbf{0} \tag{49}$$

Therefore, (46) holds. Combining this result with the update rule in (27) implies the assumption (47).

Similar to [30], [34], the inequality property (48) will be shown by comparing the Taylor series expansion of $\overline{F}(\mathbf{w})$,

$$\overline{F}(\mathbf{w}) = \overline{F}(\mathbf{w}_t) + (\mathbf{w} - \mathbf{w}_t)_I^T\nabla\overline{F}(\mathbf{w}_t)_I + \frac{1}{2}(\mathbf{w} - \mathbf{w}_t)^T(\mathbf{V}^T\mathbf{V} + \lambda\mathbf{L})(\mathbf{w} - \mathbf{w}_t) \tag{50}$$

with (43) for any $\mathbf{w}$ with $\mathbf{w}_{I'} = (\mathbf{w}_t)_{I'}$ we have

$$\overline{G}(\mathbf{w}, \mathbf{w}_t) - \overline{F}(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_t)_I^T(\overline{\mathbf{D}} - \mathbf{V}^T\mathbf{V} - \lambda\mathbf{L})_{II}(\mathbf{w} - \mathbf{w}_t)_I \tag{51}$$

therefore for this comparison we need to show that the matrix produced in our case $(\overline{\mathbf{D}} - \mathbf{V}^T\mathbf{V} - \lambda\mathbf{L})_{II}$ is positive definite which is equivalent to show that

$$\frac{[\mathbf{V}^T\mathbf{V}\overline{\mathbf{w}}_t + \lambda\overline{\mathbf{w}}_t\mathbf{L}^+]_k + \delta}{[\overline{w}_t]_k} \geq \mathbf{V}^T\mathbf{V} + \lambda\mathbf{L} \tag{52}$$

We have

$$[\mathbf{V}^T\mathbf{V}\overline{\mathbf{w}}_t]_k + \delta = \sum_{a \in I}([\mathbf{V}^T\mathbf{V}\overline{w}_t]_a + \delta) \geq [\overline{w}_t]_k\mathbf{V}^T\mathbf{V} \tag{53}$$

and

$$\lambda[\overline{\mathbf{w}}_t\mathbf{L}^+]_k = \lambda\sum_{a \in I}[\overline{w}_t\mathbf{L}^+]_a \geq \lambda[\overline{w}_t]_k\mathbf{L}^+ \geq \lambda[\overline{w}_t]_k(\mathbf{L}^+ - \mathbf{L}^-) = \lambda[\overline{w}_t]_k\mathbf{L} \tag{54}$$

Therefore, (52) holds and $\overline{G}(\mathbf{w}, \mathbf{w}_t) \geq \overline{F}(\mathbf{w})$ □

REFERENCES

[1] F. De la Torre, "A least-squares framework for component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1041–1055, Jun. 2012.

[2] E. Kokiopoulou, J. Chen, and Y. Saad, "Trace optimization and eigenproblems in dimension reduction methods," *Numer. Linear Algebra Appl.*, vol. 18, no. 3, pp. 565–602, 2011.

[3] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic, "Robust correlated and individual component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1665–1678, Aug. 2016.

[4] M. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic CCA for analysis of affective behavior and fusion of continuous annotations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1299–1311, Jul. 2014.

[5] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2000, pp. 556–562.

[6] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Comput.*, vol. 14, no. 4, pp. 715–770, Apr. 2002.

[7] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Trans. Neural Netw.*, vol. 17, no. 3, pp. 683–695, May 2006.

[8] I. Kotsia, S. Zafeiriou, and I. Pitas, "A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 588–595, Sep. 2007.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[10] M. Franzius, H. Sprekeler, and L. Wiskott, "Slowness and sparseness lead to place, head-direction, and spatial-view cells," *PLoS Comput. Biol.*, vol. 3, no. 8, p. e166, 2007.

[11] L. Zafeiriou, M. A. Nicolaou, S. Zafeiriou, S. Nikitidis, and M. Pantic, "Learning slow features for behaviour analysis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2840–2847.

[12] S. Liwicki, S. P. Zafeiriou, and M. Pantic, "Online kernel slow feature analysis for temporal video segmentation and tracking," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 2955–2970, Oct. 2015.

[13] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, vol. 14. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[14] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd ed. Belmont, MA, USA: Athena Scientific, 2000.

[15] F. Zhou and F. De la Torre, "Canonical time warping for alignment of human behavior," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2009, pp. 2286–2294.

[16] D. Gong and G. Medioni, "Dynamic manifold warping for view invariant action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 571–578.

[17] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic, "Robust canonical time warping for the alignment of grossly corrupted sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 540–547.

[18] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[19] P. Ekman and W. V. Friesen, *Facial Action Coding System*. Palo Alto, CA, USA: Consulting Psychologists Press, 1977.

[20] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 2, no. 2, pp. 121–132, 2004.

[21] M. Pantic, "Machine analysis of facial behaviour: Naturalistic and dynamic behaviour," *Phil. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 364, no. 1535, pp. 3505–3513, 2009.

[22] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Are you really smiling at me? spontaneous versus posed enjoyment smiles," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 525–538.

[23] H. Dibeklioğlu, F. Alnajar, A. A. Salah, and T. Gevers, "Combining facial dynamics with appearance for age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1928–1943, Jun. 2015.

[24] M. Pitermann and K. G. Munhall, "An inverse dynamics approach to face animation," *J. Acoust. Soc. Amer.*, vol. 110, no. 3, pp. 1570–1580, 2001.

[25] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Amsterdam, The Netherlands, Jul. 2005, pp. 317–321.

[26] M. F. Valstar and M. Pantic. *Mmi Facial Expression Database*. Accessed on May 2005. [Online]. Available: http://www.mmifacedb.com/
[27] H. Dibeklioglu, A. A. Salah, and T. Gevers. *UVA-NEMO Smile Database*. [Online]. Available: http://www.uva-nemo.org/
[28] I. N. Junejo, E. Dexter, P. Laptev, and I. Perez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 172–185, Jan. 2011.
[29] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL, USA: CRC Press, 2004.
[30] C. J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1589–1596, Nov. 2007.
[31] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 532–539.
[32] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[33] L. Zafeiriou, S. Nikitidis, S. Zafeiriou, and M. Pantic, "Slow features nonnegative matrix factorization for temporal data decomposition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 1430–1434.
[34] N. Takahashi and R. Hibi, "Global convergence of modified multiplicative updates for nonnegative matrix factorization," *Comput. Optim. Appl.*, vol. 57, no. 2, p. 417, 2014.

Lazaros Zafeiriou received the B.Sc. degree in automation from the Alexander Technical Educational Institute of Thessaloniki, Thessaloniki, Greece, in 2005, the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, in 2010, and the Ph.D. degree from the Department of Computing, Imperial College London, London, U.K., under the supervision of Prof. M. Pantic, in 2016. He has been a member of the IBUG Group, Imperial College London, since 2012. He is currently a Machine Learning Engineer with Aimbrain (authentication solutions), that designs algorithms for face and speech verification and liveness detection. His current research interests include statistical machine learning and deep learning, with an emphasis on deep metric learning, time-series analysis, and automatic human behavior analysis.

Yannis Panagakis received the B.Sc. degree in informatics and telecommunication from the National and Kapodistrian University of Athens, Greece, and the M.Sc. and Ph.D. degrees from the Department of Informatics, Aristotle University of Thessaloniki. He is currently a Lecturer (equivalent to Assistant Professor) in computer science with Middlesex University, London, and a Research Fellow with the Department of Computing, Imperial College London. His research interests include machine learning, signal processing, and mathematical optimization with applications to computer vision, human behavior analysis, and music information research. He received various scholarships and awards for his studies and research, including the prestigious Marie-Curie Fellowship in 2013. He is the Workshops Chair in BMVC 2017. He currently serves as an Associate Editor of the *Image and Vision Computing Journal*. His work has been featured in top venues in the field, such as the IEEE T-PAMI, the TIP, the IJCV, the CVPR, and the ICCV.

Maja Pantic is currently a Professor in affective and behavioral computing with the Department of Computing, Imperial College London, U.K., and also with the Department of Computer Science, University of Twente, The Netherlands. She has received various awards for her work on automatic analysis of human behavior, including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She currently serves as the Editor-in-Chief of the *Image and Vision Computing Journal* and an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.

Stefanos Zafeiriou is currently a Senior Lecturer in pattern recognition/statistical machine learning for computer vision with the Department of Computing, Imperial College London, London, U.K, and a Distinguishing Research Fellow with the University of Oulu, under Finish Distinguishing Professor Programme. He was a recipient of the Prestigious Junior Research Fellowships from Imperial College London in 2011 to start his own independent research group. He was a recipient of the President's Medal for Excellence in Research Supervision for 2016, and various awards during his Doctoral and Post-Doctoral studies. He is the General Chair of BMVC 2017. He currently serves as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS and the *Image and Vision Computing Journal*. He has been a Guest Editor of over six journal special issues and co-organized over nine workshops/special sessions on face analysis topics in top venues, such as CVPR/FG/ICCV/ECCV, including two very successfully challenges run in ICCV13 and ICCV15 on facial landmark localization/tracking. He has over 2800 citations to his work, h-index 27.