Joint Unsupervised Face Alignment and Behaviour Analysis*

Lazaros Zafeiriou, Epameinondas Antonakos,
Stefanos Zafeiriou, and Maja Pantic

Computing Department, Imperial College London, UK
{l.zafeiriou12,e.antonakos,s.zafeiriou,m.pantic}@imperial.ac.uk

Abstract. The predominant strategy for facial expressions analysis and temporal analysis of facial events is the following: a generic facial landmarks tracker, usually trained on thousands of carefully annotated examples, is applied to track the landmark points, and then analysis is performed using mostly the shape and more rarely the facial texture. This paper challenges the above framework by showing that it is feasible to perform joint landmarks localization (i.e. spatial alignment) and temporal analysis of behavioural sequence with the use of a simple face detector and a simple shape model. To do so, we propose a new component analysis technique, which we call Autoregressive Component Analysis (ARCA), and we show how the parameters of a motion model can be jointly retrieved. The method does not require the use of any sophisticated landmark tracking methodology and simply employs pixel intensities for the texture representation.

Keywords: Face alignment, time series alignment, slow feature analysis.

1 Introduction

The analysis of facial Action Units (FAUs) and expressions are important tasks in Computer Vision and Human-Computer Interaction, which have accumulated great research effort [1]. The standard approach is the application of a robust facial tracker for the facial landmark points localization and then the application of an analysis technique. The tracker can be either generic or person-specific, depending on the task and the available annotations [2,3]. On the one hand, methodologies that show exceptional performance in generic facial tracking have been recently proposed [4,5,6], capitalizing on the abundance of databases with thousands of annotated facial images in both controlled [7] and uncontrolled conditions [8,9]. On the other hand, the person-specific tracker framework requires manual annotation of a number of frames from a person's video sequence. The manual annotation of images, which is required by such methods, is a very time consuming, expensive and labour intensive procedure. Furthermore, the expressions and FAUs analysis is performed using mainly the geometric displacement

* Electronic supplementary material -Supplementary material is available in the online version of this chapter at http://dx.doi.org/10.1007/978-3-319-10593-2_12. Videos can also be accessed at http://www.springerimages.com/videos/978-3-319-10592-5

D. Fleet et al. (Eds.): ECCV 2014, Part IV, LNCS 8692, pp. 167–183, 2014.

of facial shape points [10,11,12] and secondarily the facial texture in the form of hand-crafted features, i.e. Local Binary Patterns (LBPs) and SIFT features ([13],[14]). Finally, when it comes to temporal alignment of facial events, the tracked facial landmarks are aligned after being tracked [3,15,16], usually by the application of a person specific tracker.

In this paper we take a radically different direction. We propose a methodology that can be used to perform joint automatic facial landmarks localization and discovery of features that can be used for analysis of temporal events (e.g. analysis of FAU dynamics). To do so, we start by formulating a special undirected Gaussian Hidden Markov Random Field. The GHMRF is a generative model which jointly describes (i.e. generates) the data and also captures temporal dependencies by incorporating an autoregressive chain [17] in the latent space. We show how a novel deterministic component analysis, which we coin Autoregressive Component Analysis (ARCA), can be formulated. We further show how a motion model can be incorporated in ARCA. Our methodology has been motivated by the success of joint alignment and low-rank matrix recovery in person specific scenarios [18,19,20] as well as previous works on parametrized component analysis [21,22]. But our method is radically different to [18], since (1) it extracts latent features rather than image reconstructions, (2) it incorporates a non-rigid motion model guided by a shape model rather than rigid motion used in [18][1] and (3) it incorporates time dependencies. Furthermore, our method is radically different to [24] and [20] which are based on trained models of appearance and require annotations of hundreds of images to allow good generalization.

By extending such methodologies in order to take into account the correlations between sequences that depict the same facial event (i.e. FAU), we show that the extracted features can be used to perform temporal alignment. Moreover, we show that the proposed method achieves successful results even though it does not utilize any robust feature-based representation of the appearance (e.g. HOG, SIFT) as usually done in the literature, but it is instead applied on the pixel intensities. Summarizing, the contributions of the paper are:

- We propose a novel component analysis which can perform joint reconstruction and extraction of a latent space with first order Markov dependencies. Hence, the proposed component analysis can be used for joint construction of a deformable model and extraction of smooth features for event analysis
- 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.

We would like to note here that the extracted features are more suitable for unsupervised segmentation of behaviour analysis of behaviour dynamics, as well as temporal alignment rather than recognition of expression and/or action units.

[1] Recently it has been empirically shown that [20,23], due to the presence of an non-regularised low-rank term, the method in [18] fails in case of non-rigid motion.

The only prerequisites of the proposed method are the presence of (1) a simple bounding box face detector and (2) a shape model, by means of a Point Distribution Model (PDM), of the facial landmarks that we want to detect. The face detector can be as simple as the Viola-Jones object detector [25] which can return only the true positive detection of a face's bounding box. Such detectors are widely and successfully used. For example, the newest versions of Matlab have incorporated a training procedure of Viola-Jones. Additionally, such detectors are also widely employed in commercial products (e.g. even the cheapest digital camera has a robust face detector). Besides, the annotations that are needed to train such a detector can be acquired very quickly, since only a bounding box containing the image's face is required. Other detectors that can be used are efficient subwindow search [26] and deformable part-based models [27,28,24]. The statistical shape model of facial landmark points can be built easily using a small number of facial shapes. Around 50 shapes of images from the internet are sufficient in order to build a descriptive shape model that can generate multiple facial expressions and their annotation takes less than 4 hours. Finally, there are unsupervised techniques to learn the shape model directly from images [29,30].

2 Method

2.1 Definitions and Prerequisites

We assume that we have a set of facial shapes and a crude face detector, such as Viola-Jones [25]. We denote a facial shape as a $2L_S \times 1$ vector $\mathbf{s} = [x_1, y_1, \ldots, x_{L_S}, y_{L_S}]^T$, where $(x_i, y_i),\ i = 1, \ldots, L_S$ are the coordinates of the L_S landmark points. The PDM shape model consists of an orthonormal basis $\mathbf{U}_S \in \mathbb{R}^{2L_S \times N_S}$ of N_S eigenvectors and the mean shape $\bar{\mathbf{s}}$, which are derived from the facial shapes in our disposal. Note that the first four eigenvectors correspond to the global similarity transform that controls the face's rotation, scaling and translation. A new shape instance is generated as a linear combination of the eigenvectors weighted by the parameters $\mathbf{p} = [p_1, \ldots, p_{N_S}]^T$, thus $\mathbf{s_p} = \bar{\mathbf{s}} + \mathbf{U}_S\mathbf{p}$. Moreover, let us denote a motion model as the warp function $\mathcal{W}(\mathbf{x}, \mathbf{p})$, which maps each point within the mean (reference) shape ($\mathbf{x} \in \bar{\mathbf{s}}$) to its corresponding location in a shape instance. We employ the Piecewise Affine Warp which performs the mapping based on the barycentric coordinates of the corresponding triangles between the source and target shapes that are extracted using Delaunay triangulation. In the rest of the paper, we will denote the warp function as $\mathcal{W}(\mathbf{p})$ for simplicity.

2.2 Autoregressive Component Analysis with Spatial Alignment

In this section we propose a deterministic component analysis based on an Autoregressive (AR) statistical model. In particular, we start by formulating a probabilistic generative model which (1) captures time-variant latent features and (2) explains data generation. Hence, it can be used for joint extraction of

latent features which capture time dependencies and, in the same time, as a linear statistical model suitable for deformable model construction.

Assume that we have a time-variant, multi-dimensional input signal, e.g. a video sequence of N frames, denoted in vectorized form as $\mathbf{x}_i \in \mathbb{R}^F$, $i = 1, \ldots, N$, which shows a person that performs a facial expression or FAU. The frames' appearance is based on pixel intensities. We denote as $\mathbf{X} \in \mathbb{R}^{F \times N}$ the matrix that has these vectorized frames as its columns, thus $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$.

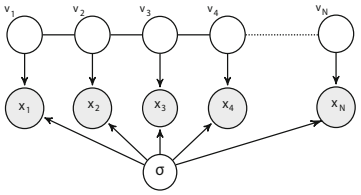


Fig. 1. Graphical model of an Autoregressive process

We assume a generative model in the form of $\mathbf{x}_i = \mathbf{U}\mathbf{v}_i + \mathbf{e}_i$, where $\mathbf{U} \in \mathbb{R}^{F \times K}$ is a subspace of K basis ($K < min(F, N)$). We also assume that $\mathbf{e}_i$ follows a zero mean Gaussian distribution with $\sigma^2\mathbf{I}$ covariance matrix, thus $\mathbf{e}_i \sim \mathcal{N}(\mathbf{e}_i|\mathbf{0}, \sigma^2\mathbf{I})$. Furthermore, in order to capture the time-variant correlations of the signals, we assume an AR model for the latent space $\mathbf{v}_i$ as $\mathbf{v}_i|\mathbf{v}_{i-1}, \ldots, \mathbf{v}_1 \sim \mathcal{N}(\mathbf{v}_i|\phi\mathbf{v}_{i-1}, \mathbf{I})$ with $\mathbf{v}_1 \sim \mathcal{N}(\mathbf{v}_1|\mathbf{0}, (1-\phi^2)^{-1})$. The graphical model of the AR model is shown in Fig. 1. That is, assume the matrix $\mathbf{V}$ of the latent features with columns $\mathbf{v}_i$, i.e. $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N] \in \mathbb{R}^{K \times N}$, and its K rows denoted by $\tilde{\mathbf{v}}_j$ with size $N \times 1$. Each row is an AR model, which is a special case of a Gaussian Markov Random Field (GMRF) [17]

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

with the tridiagonal precision matrix $\mathbf{L} \in \mathbb{R}^{N \times N}$ given by

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

The probability for all the rows of matrix $\mathbf{V}$ can be written as

$$p(\mathbf{V}|\mathbf{L}) = \prod_{j=1}^{K} p(\tilde{\mathbf{v}}_j|\mathbf{L}) = \frac{|\mathbf{L}|^N}{\sqrt{(2\pi)^{KN}}} e^{-\frac{1}{2}\sum_{j=1}^{N}(\tilde{\mathbf{v}}_\mathbf{j})^T \mathbf{L} \tilde{\mathbf{v}}_j} = \frac{|\mathbf{L}|^N}{\sqrt{(2\pi)^{KN}}} e^{-\frac{1}{2}\,\mathrm{tr}[\mathbf{V}\mathbf{L}\mathbf{V}^T]} \tag{3}$$

where tr[.] denotes the matrix trace operator. Hence, according to Fig. 1, the factorization of the joint likelihood of $\mathbf{X}, \mathbf{V}$ given σ^2, $\mathbf{L}$ and $\mathbf{U}$ has the form

$$\begin{aligned} p(\mathbf{X}, \mathbf{V}|\mathbf{L}, \mathbf{U}, \sigma^2) &= p(\mathbf{X}|\mathbf{V}, \mathbf{U}, \sigma^2) p(\mathbf{V}|\mathbf{L}) \\ &= \prod_{i=1}^{N} p(\tilde{\mathbf{x}}_i|\mathbf{v}_i, \mathbf{L}, \sigma^2) p(\mathbf{V}|\mathbf{L}) \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^{NF}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{U}\mathbf{v}_i)^T(\mathbf{x}_i - \mathbf{U}\mathbf{v}_i)} \frac{|\mathbf{L}|^N}{\sqrt{(2\pi)^{KN}}} e^{-\frac{1}{2}\,\mathrm{tr}[\mathbf{V}\mathbf{L}\mathbf{V}^T]} \\ &= \frac{|\mathbf{L}|^N}{\sqrt{(\sigma^2)^{NF}(2\pi)^{N(K+F)}}} e^{-\frac{1}{2}(\frac{1}{\sigma^2}||\mathbf{X} - \mathbf{U}\mathbf{V}||_F^2 + \mathrm{tr}[\mathbf{V}\mathbf{L}\mathbf{V}^T])} \end{aligned} \tag{4}$$

where $||.||_F$ denotes the matrix Frobenius norm. Taking the logarithm of the above joint probability, we get a cost function with regards to $\mathbf{U}, \mathbf{V}$

$$\begin{aligned} g(\mathbf{U}, \mathbf{V}) &= \ln p(\mathbf{X}, \mathbf{V}|\mathbf{L}, \mathbf{U}, \sigma^2) \\ &\propto -||\mathbf{X} - \mathbf{U}\mathbf{V}||_F^2 - \lambda \mathrm{tr}[\mathbf{V}\mathbf{L}\mathbf{V}^T] + \mathrm{const} \end{aligned} \tag{5}$$

For simplicity, we set $\phi = 0.9$ and the variance $\sigma^2 = \frac{1}{\lambda} = 0.1$, where $\lambda \geq 0$ is a regularization parameter that controls the smoothness of the method that is used to compute the matrix $\mathbf{L}$. The first term $||\mathbf{X} - \mathbf{U}\mathbf{V}||_F^2$ of Eq. 5 measures how well the data can be reconstructed from the loading matrix $\mathbf{U}$ and the latent space weights $\mathbf{V}$, while the second term $\mathrm{tr}[\mathbf{V}\mathbf{L}\mathbf{V}^T]$ is a smoothing constraint over the latent space to model the undirected temporal dependencies. If we impose further orthogonality constraints on $\mathbf{U}$, we get

$$\min_{\mathbf{U},\mathbf{V}} \quad f(\mathbf{U}, \mathbf{V}) = ||\mathbf{X} - \mathbf{U}\mathbf{V}||_F^2 + \lambda \mathrm{tr}[\mathbf{V}\mathbf{L}\mathbf{V}^T] \quad \text{s.t.} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I} \tag{6}$$

where $\mathbf{I}$ denotes the identity matrix. In order to get meaningful results that explain the actual variations of images and not the variations due to misalignment, as done in all component analysis techniques [31], solving Eq. 6 requires to provide perfectly aligned images achived through manual annotations.

In this paper, we propose to take a radically different approach and jointly find the components $\mathbf{U}$, the time-variant latent space $\mathbf{V}$ and a set of parameters that align the images into a common frame, defined by the mean shape $\bar{\mathbf{s}}$. In order to do so, we introduce warp parameters on the data matrix $\mathbf{X}$. The warping of each video frame in the mean (reference) shape, given a shape estimate of the frame's displayed face ($\{\mathbf{s}_i\},\ i = 1, \ldots, N$), returns N appearance vectors $\{\mathbf{x}_i(\mathcal{W}(\mathbf{p}_i))\},\ \forall i = 1, \ldots, N$ of size $F \times 1$, where F is the number of pixels that lie inside the mean shape. We denote as

$$\mathbf{X}(\mathcal{W}(\mathbf{P})) = [\mathbf{x}_1(\mathcal{W}(\mathbf{p}_1)), \ldots, \mathbf{x}_N(\mathcal{W}(\mathbf{p}_N))] \tag{7}$$

the $F \times N$ time-varying input data matrix that consists of the warped frames' vectors, where

$$\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_N]$$

is the matrix of the shape parameters of each frame. The cost function of Eq. 6 now becomes

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V},\mathbf{P}} \quad & f(\mathbf{U}, \mathbf{V}, \mathbf{P}) = \|\mathbf{X}(\mathcal{W}(\mathbf{P})) - \mathbf{U}\mathbf{V}\|_F^2 + \lambda\, \mathrm{tr}[\mathbf{V}\mathbf{L}\mathbf{V}^T] \\ \text{s.t.} \quad & \mathbf{U}^T\mathbf{U} = \mathbf{I} \end{aligned} \tag{8}$$

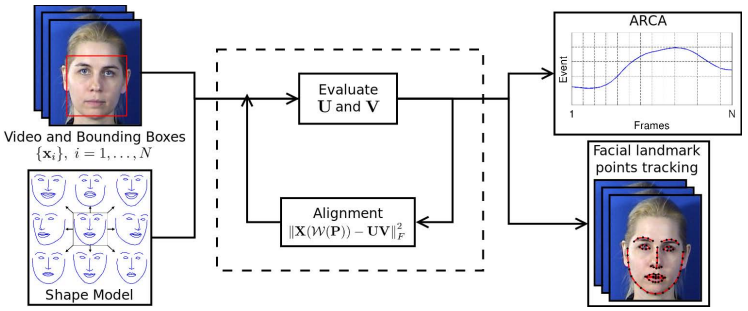


Fig. 2. Method overview. Given a video sequence with the corresponding bounding boxes and a shape model, the method performs joint facial landmarks localization and spatio-temporal facial behaviour analysis.

We solve the minimization of Eq. 8 in an alternating manner, as shown in Fig. 2. In brief, the method iteratively solves for matrices $\mathbf{U}$ and $\mathbf{V}$ based on the current estimate of the warped vectors $\mathbf{X}(\mathcal{W}(\mathbf{P}))$ and then re-estimates the shape parameters $\mathbf{P}$ of the sequence's frames. The initial shapes are estimated by applying a similarity transform on the mean shape $\bar{\mathbf{s}}$ to confront fit within the boundaries of each frame's bounding box. This means that the initial shape parameters are equal to zero ($\mathbf{p}_i = \mathbf{0},\ \forall i = 1, \ldots, N$). Consequently, the optimizaton is solved in the following two steps:

Fix P and Minimize with Respect to {U, 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 $\mathbf{U}$. Given $\mathbf{V}$ the optimization problem with regards to $\mathbf{U}$ is given by

$$\mathbf{U}_o = \underset{\mathbf{U}}{\operatorname{argmin}}\, f(\mathbf{U}) = \|\mathbf{X}(\mathcal{W}(\mathbf{P})) - \mathbf{UV}\|_F^2 \quad \text{s.t} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}. \tag{9}$$

The solution of the above optimization problem is given by the skinny singular value decomposition (SSVD) of $\mathbf{X}(\mathcal{W}(\mathbf{P}))\mathbf{V}^T$ [32]. That is, if the SVD of $\mathbf{X}(\mathcal{W}(\mathbf{P}))\mathbf{V}^T = \mathbf{RSM}^T$, then

$$\mathbf{U} = \mathbf{RM}^T. \tag{10}$$

Updating $\mathbf{V}$. Given $\mathbf{U}$ the optimization problem with regards to $\mathbf{V}$ is given by

$$\mathbf{V}_o = \underset{\mathbf{V}}{\operatorname{argmin}}\, f(\mathbf{V}) = \|\mathbf{X}(\mathcal{W}(\mathbf{P})) - \mathbf{UV}\|_F^2 + \lambda \operatorname{tr}[\mathbf{VLV}^T] \tag{11}$$

which gives the update

$$\mathbf{V} = \mathbf{U}^T\mathbf{X}(\mathcal{W}(\mathbf{P}))(\mathbf{I} - \lambda\mathbf{L})^{-1} \tag{12}$$

Fix {U, 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 $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_N]$ 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 [33]. The cost function of this step can be written as

$$\min_{\mathbf{P}} \|\mathbf{X}(\mathcal{W}(\mathbf{P})) - \mathbf{UV}\|_F^2 = \min_{\{\mathbf{p}_i\},\ i=1,\ldots,N} \sum_{i=1}^{N} \|\mathbf{x}_i(\mathcal{W}(\mathbf{p}_i)) - \mathbf{U}\mathbf{v}_i\|_2^2 \tag{13}$$

where $\mathbf{v}_i,\ \forall i = 1, \ldots, N$ denotes the i^{th} column of the matrix $\mathbf{V}$. We solve the problem of Eq. 13 by minimizing for each frame separately, as

$$\min_{\mathbf{p}_i} \|\mathbf{x_i}(\mathcal{W}(\mathbf{p}_i)) - \mathbf{y}_i\|_2^2,\ i = 1, \ldots, N \tag{14}$$

where $\mathbf{y}_i = \mathbf{U}\mathbf{v}_i$ denotes the template corresponding to each frame. Within the IC optimization technique, an incremental warp is introduced on the part of the template of Eq. 14, thus the aim is to minimize

$$\min_{\Delta\mathbf{p}_i} \|\mathbf{x_i}(\mathcal{W}(\mathbf{p}_i)) - \mathbf{y}_i(\mathcal{W}(\Delta\mathbf{p}_i))\|_2^2 \tag{15}$$

with respect to $\Delta\mathbf{p}_i$. Then, at each iteration, a compositional update rule is applied on the shape parameters, as

$$\mathcal{W}(\mathbf{p}_i) \leftarrow \mathcal{W}(\mathbf{p}_i) \circ \mathcal{W}(\Delta\mathbf{p}_i)^{-1}$$

The solution of Eq. 15 is derived by taking the first-order Taylor expansion of the template term around $\Delta\mathbf{p}_i = \mathbf{0}$ and using the identity property of the warp function ($\mathcal{W}(\mathbf{x}, \mathbf{0}) = \mathbf{x}$), as $\mathbf{y}_i(\mathcal{W}(\Delta\mathbf{p}_i)) \approx \mathbf{y}_i + \mathbf{J}_{\mathbf{y}_i}|_{\mathbf{p}=\mathbf{0}}\Delta\mathbf{p}_i$, where $\mathbf{J}_{\mathbf{y}_i}|_{\mathbf{p}=\mathbf{0}} = \nabla\mathbf{y}_i \left.\frac{\partial\mathcal{W}}{\partial\mathbf{p}}\right|_{\mathbf{p}=\mathbf{0}}$ is the template Jacobian that consists of the template gradient and the warp Jacobian evaluated at $\mathbf{p} = \mathbf{0}$. Substituting this linearization to Eq. 15, the solution is given by

$$\Delta\mathbf{p}_i = \mathbf{H}^{-1}\mathbf{J}_{\mathbf{y}_i}^T|_{\mathbf{p}=\mathbf{0}}\,[\mathbf{x_i}(\mathcal{W}(\mathbf{p}_i)) - \mathbf{y}_i]$$

where $\mathbf{H} = \mathbf{J}_{\mathbf{y}_i}^T|_{\mathbf{p}=\mathbf{0}}\mathbf{J}_{\mathbf{y}_i}|_{\mathbf{p}=\mathbf{0}}$ is the Gauss-Newton approximation of the Hessian matrix. Note that since the gradient is always computed at the template (reference frame), the warp Jacobian and the Hessian matrix inverse remain constant, which results in a small computational cost.

3 Comparison with State-of-the-Art Component Analysis Techniques

Even though component analysis is a very well-studied research field including very popular methodologies such as Principal Component Analysis (PCA)

[34], Linear Discriminant Analysis (LDA)[35] and Graph Embedding techniques [36,37], there is very limited work on deterministic component analysis techniques for discovering latent spaces that capture time dependencies[2]. One such component analysis is the so-called Slow Feature Analysis (SFA) [39], which aims to identify the most slowly varying features from rapidly temporal varying signals. More formally, given an F-dimensional time-varying input sequence, SFA seeks to determine appropriate projection bases stored in the columns of matrix $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$, that in the low dimensional space minimize the variance of the approximated first order time derivative of the latent variables $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N] = \mathbf{U}^T\mathbf{X}$, subject to zero mean, unit covariance and decorrelation constraints

$$\min_{\mathbf{U}} \text{tr}[\mathbf{U}^T\dot{\mathbf{X}}\dot{\mathbf{X}}^T\mathbf{U}] \quad \text{s.t.} \quad \mathbf{V1} = \mathbf{0},\ \mathbf{U}^T\mathbf{X}\mathbf{X}^T\mathbf{U} = \mathbf{I} \tag{16}$$

where $\mathbf{1}$ is a $N \times 1$ vector with all its elements equal to $\frac{1}{N}$. The matrix $\dot{\mathbf{X}} \in \mathbb{R}^{F\times(N-1)}$ approximates the first order time derivative of $\mathbf{X}$, evaluated by taking the temporal differences between successive sample observations, as

$$\dot{\mathbf{X}} = [\mathbf{x}_2 - \mathbf{x}_1, \mathbf{x}_3 - \mathbf{x}_2, \ldots, \mathbf{x}_N - \mathbf{x}_{N-1}] = \mathbf{X}\mathbf{Q} \tag{17}$$

where $\mathbf{Q}$ is an $N \times (N-1)$ matrix with elements $q_{i,i} = -1$, $q_{i+1,i} = 1$ and 0 elsewhere. The optimal $\mathbf{U}$ from Eq. 16 is given as the eigenvectors of $[\mathbf{X}\mathbf{X}^T]^{-1}[\dot{\mathbf{X}}\dot{\mathbf{X}}^T]$ that correspond to the smallest eigenvalues. We should note that since SFA introduces an ordering to the derived latent variables sorted by the temporal slowness, the smallest eigenvectors correspond to the slowest varying features. In the following, we show that an orthogonal variant of SFA can be derived as a special case of ARCA. In particular, assuming a uniform prior for $p(\mathbf{v}_1)$ (i.e. $\phi = 1$), then the precision matrix $\mathbf{L}$ can be decomposed as $\mathbf{L} = \mathbf{Q}\mathbf{Q}^T$ and by substituting $\mathbf{V} = \mathbf{U}^T\mathbf{X}$ in Eq. 6, the optimization problem can be reformulated as

$$\begin{aligned}
\min_{\mathbf{U}} f(\mathbf{U}) =& ||\mathbf{X} - \mathbf{U}\mathbf{U}^T\mathbf{X}||_F^2 + \lambda \text{tr}[\mathbf{U}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{U}] \\
=& \text{tr}[\mathbf{X}^T\mathbf{X}] - \text{tr}[\mathbf{U}^T\mathbf{X}\mathbf{X}^T\mathbf{U}] + \lambda \text{tr}[\mathbf{U}^T\mathbf{X}\mathbf{Q}\mathbf{Q}^T\mathbf{X}^T\mathbf{U}] \\
=& -\text{tr}[\mathbf{U}^T\mathbf{X}\mathbf{X}^T\mathbf{U}] + \lambda \text{tr}[\mathbf{U}^T\dot{\mathbf{X}}\dot{\mathbf{X}}^T\mathbf{U}] + \text{const} \\
=& \text{tr}[\mathbf{U}^T(\lambda\dot{\mathbf{X}}\dot{\mathbf{X}}^T - \mathbf{X}\mathbf{X}^T)\mathbf{U}] + \text{const} \quad \text{s.t.} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}.
\end{aligned} \tag{18}$$

where $\mathbf{U}$ stores the K non-zero eigenvectors that correspond to the K smallest eigenvalues of $\lambda\dot{\mathbf{X}}\dot{\mathbf{X}}^T - \mathbf{X}\mathbf{X}^T$. Hence, the optimization problem of Eq. 18 gives a similar result, but imposes an extra orthogonality on $\mathbf{U}$.

4 Experiments

The experiments aim to demonstrate that the proposed unsupervised procedure is able to locate landmarks so as to perform image alignment and in the same

[2] There is very rich literature about Gaussian Linear Dynamical Models, i.e. Kalman filters [38], but this is a totally different way of modelling time series, which in principle cannot be easily combined with spatial warping techniques as the ARCA.

time extract latent features that can reveal the dynamics of facial behaviour, directly from image intensities. The gold standard in unsupervised behaviour analysis is (a) to track facial landmark points and (b) use their motion to perform analysis. For example in [2,10] person specific trackers were used, which require manual annotation, and in [15,3] a generic tracker was employed followed by a manual correction step. The goals of the experiments are two fold: (1) to show that the method can correctly track landmarks from a crude face detector and (2) to show that the extracted features can represent the dynamics of the behaviour. To do so, we use two databases: MMI[40,41] that has posed FAUs and UvA-Nemo Smile (UNS) [42] that displays more complex spontaneous behaviour. MMI consists of more than 400 videos annotated in terms of FAUs and the temporal phases in which a subject performs one or more FAUs. We use 61 of those videos, which are the ones that we manually annotated with 68 landmarks in order to compare. UvA-Nemo Smile database is a large-scale database having more than 1000 smile videos (597 spontaneous and 643 posed) from 400 subjects. Similarly to the MMI database, we conduct experiments on 25 videos with spontaneous smiles, which we manually annotated in terms of the smile's temporal phases and the 68 facial landmark points.

In ARCA we employ a shape model trained on 50 shapes of Multi-Pie database [7], annotated with the same $L_S = 68$ landmark configuration. The model consists of $N_S = 15$ eigenvectors and the mean (reference) shape has a resolution of 169×171, thus the dimensionality of our data matrix is $F = 28899$. Moreover, the faces' bounding boxes of all the videos are detected using the Viola-Jones object detection algorithm [25]. Finally, the proposed method is applied using 5 global iterations. In Section 4.1 we show results on the spatio-temporal behaviour analysis of the videos and in Section 4.3 we present the facial landmarks localization performance. Throughout the experiments, we set the regularization parameter that controls the smoothness of the proposed method equal to $\lambda = 10$ and we limit the number of extracted basis to $K = 30$.

4.1 Spatio-temporal Behaviour Analysis Results in MMI Database

In this section we provide experimental results for the task of unsupervised facial behaviour analysis. Specifically, we investigate how accurately the proposed method can capture the transitions between the temporal phases during the activation of various FAUs and compare against SFA. The temporal phases of a performed FAU are: (1) *Neutral* when the face is relaxed, (2) *Onset* when the action initiates, (3) *Apex* when the muscles reach the peak intensity and (4) *Offset* when the muscles begin to relax. The performance of the methods is evaluated by comparing the slowest varying features extracted by both methods with the ground truth annotations. To identify which of the extracted feature corresponds to the most slowly varying one we computed the first order time derivative for each obtained latent variable and keep the one with minimum: $\mathbf{v}_i^T \mathbf{L} \mathbf{v}_i$. For comparison we apply SFA on the ground truth shape [3]. More precisely, we measure

[3] The result of SFA on texture did not capture the dynamics.

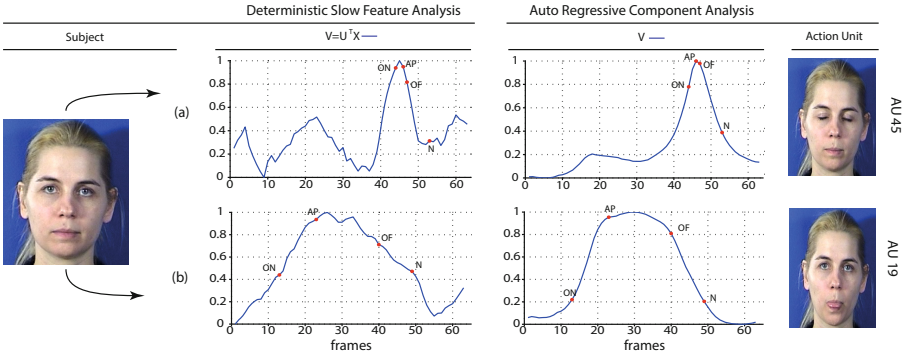


Fig. 3. Application of SFA and ARCA on a video from MMI database displaying a subject performing: (a) Blink (AU 45) and (b) Tongue Show (AU 19). The red marks indicate the ground truth moments at which the FAU's temporal phases change (ON - neutral to onset, AP - onset to apex, OF - apex to offset, N - offset to neutral).

the similarity between the ground truth and the extracted features by monitoring the alignment cost using the dynamic time warping (DTW) algorithm. Therefore, a low measured cost means that the FAUs transitions are captured more accurately by the extracted feature.

Figure 3 shows the performance of the proposed method against SFA in terms of capturing the FAU temporal phases from a subject that performs two AUs in the same video sequence. More specifically, Figs. 3(a) and 3(b) show the results obtained when the subject performs AU45 (i.e. blink) and AU19 (i.e tongue show) respectively. In each plot the red marks correspond to the ground truth points at which the FAU's temporal phase changes. The graphs of both sequences indicate that the proposed method outperforms the SFA algorithm since it detects the dynamics of the FAU more accurately and captures the temporal phases more smoothly.

Figure 4 shows the error between the extracted features and the ground truth annotations for the MMI database's videos with the application of both the ARCA and SFA. More precisely, Fig. 4(a) shows the error from 53 videos in which the subject performs mouth-related FAUs, while Fig. 4(b) shows the error from 35 videos in which the subject performs eyes-related FAUs. Table 1 summarizes these results for each temporal phase separately. The presented results indicate that the proposed method significantly outperforms SFA on the unsupervised detection of the temporal phases of FAUs, almost in all temporal phases and for all relevant regions of the face.

Next we test the ability of the ARCA method to provide low dimensional texture features that can be used for temporal alignment of behaviour. To do so, we combine the extracted features from ARCA with DTW. We compare this method with Canonical Time Warping (CTW) [2], which jointly discovers low dimensional features that can be used for temporal alignment of sequences. For CTW, we used the textures aligned using the ground truth shapes. In the

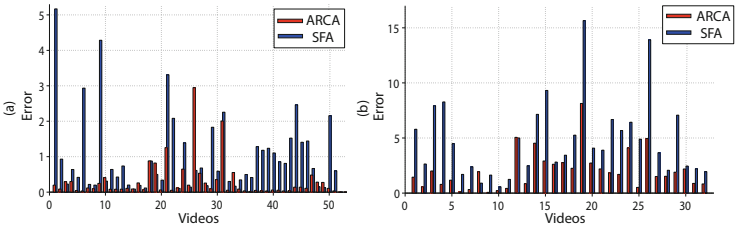


Fig. 4. Total error between the extracted features and ground truth annotations on the MMI database. The plots compare the performance of the proposed method and SFA with: (a) Mouth-related AUs (b) Eyes-related AUs.

Table 1. Error between the extracted features and ground truth annotations for each temporal phase on the MMI database. The results compare the performance of the fully automatic ARCA method against SFA on ground truth shape.

	Neutral			Onset			Apex			Offset		
Method	Mouth	Eyes	Brows	Mouth	Eyes	Brows	Mouth	Eyes	Brows	Mouth	Eyes	Brows
ARCA	**0.341**	**2.299**	**0.388**	**0.215**	**0.104**	**0.053**	**0.516**	**0.252**	**0.266**	**0.2534**	1.298	**0.2638**
SFA	1.054	3.943	2.154	0.675	0.329	0.277	2.541	2.889	0.705	0.506	**1.076**	1.084

example shown in Fig. 5 two different subjects perform AU10 (Upper Lip Raiser) in different moments. As can be observed in Fig. 5(b), ARCA+DTW was able to align accurately all the temporal phases, while the low dimensional features provided by CTW were not able to align the sequences, as indicated by its respective alignment path 5(c) solid line. For further alignment examples, please see the supplementary material. Fig. 5(d) shows several frames illustrating the alignment.

4.2 Behaviour Analysis of Spontaneous Smiles in UVS Database

As it is widely shown spontaneous behaviour differs greatly to posed behaviour both in duration and dynamics [1]. In particular, in spontaneous behavior it is very often that we do not have a single smooth transition but we have many valleys and plateaus. In order to evaluate whether the proposed methodology can capture this complex transitions we used the spontaneous smiles of UNS database.

Figure 6 shows an example in which the subject performs an FAU with many transitions. This means that the performed FAU has more than one onset and apex phases. During the first apex phase (frames 24 to 94) the subject is smiling with a normal intensity. However, the smile intensifies during frames 94 to 102 and reaches its second peak at frame 103. As can be seen in the graph, the proposed method manages to capture all the transitions of the temporal phases more accurately and smoothly compared to SFA. Moreover, Table 2 summarizes the results on all UNS database videos. Specifically, it reports the mean error

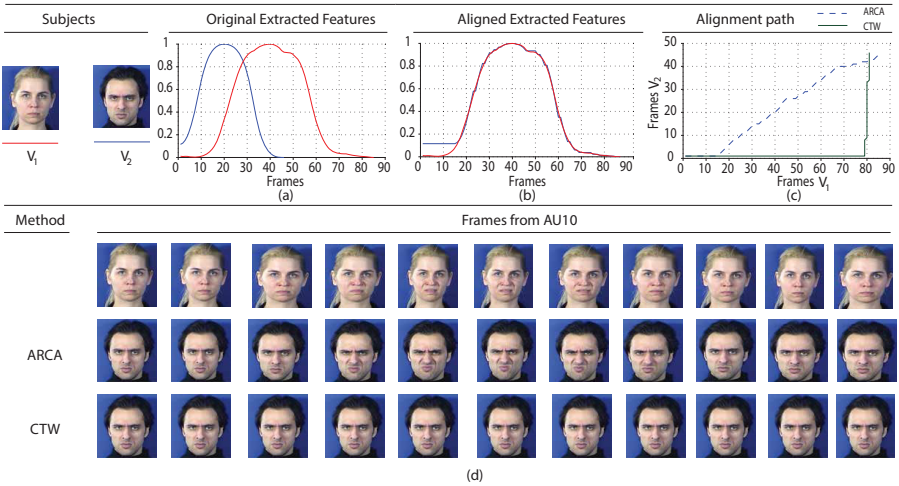


Fig. 5. Aligning the AU10 performed by two different subjects. (a)Original features (b)Aligned features (c) Alignment path. (d) Frames detected form the ARCA method (second row) and CTW method (third row).

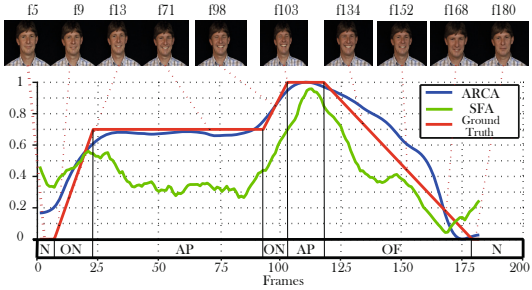


Fig. 6. Comparison of ARCA (blue) and SFA (green) with the annotated ground truth (red) on a spontaneous video sequence from UNS database. The subject performs an FAU with multi-temporal phases (ON-Onset, AP-apex, OF-offset, N-neutral).

of each temporal phase along with the overall error of the whole performed FAU. Similar to the MMI experiments, the results show that ARCA significantly outperforms SFA on the unsupervised detection of the multi temporal phases of AUs in all temporal phases.

4.3 Landmark Points Localization Results

In this section we present experimental results for the task of automatic facial landmarks localization. We evaluate the error between an estimated shape and the ground truth with the point-to-point RMSE measure normalized with respect to the face's size. Specifically, denoting as $\mathbf{s}^f$ and $\mathbf{s}^g$ the fitted and ground

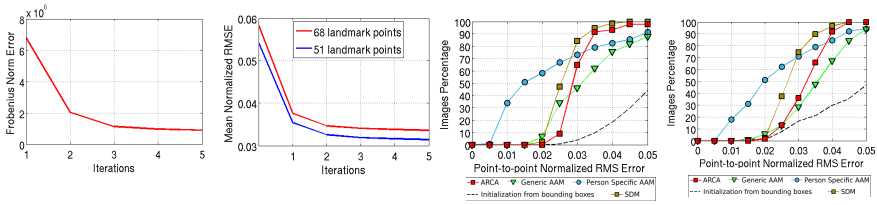
Table 2. Error between the extracted features and ground truth annotations for each temporal phase on the UNS database. The results compare the performance of the fully automatic ARCA method against SFA on ground truth shape.

Method	Neutral	Onset	Apex	Offset	Overall
ARCA	**0.147**	**0.087**	**0.791**	**0.050**	**0.1524**
SFA	2.068	0.610	8.250	0.497	2.081

truth shapes respectively, the normalized RMSE between them is $RMSE = \frac{\sum_{i=1}^{L_S} \sqrt{(x_i^f - x_i^g)^2 + (y_i^f - y_i^g)^2}}{L_S d}$ where $d = (\max_x \mathbf{s}^g - \min_x \mathbf{s}^g + \max_y \mathbf{s}^g - \min_y \mathbf{s}^g)/2$ is face's size.

Figures 7a and 7b provide a proof that the cost function converges. Specifically, Fig. 7a shows the evolution of the mean cost function error of Eq. 13 over all MMI database's videos with respect to the iterations. As can be seen, the error monotonically decreases. Additionally, Fig. 7b visualizes the evolution of the mean normalized RMSE between the fitted shapes and the ground truth annotations over all MMI database's videos with respect to the iterations. Note that the plot shows the RMSE evaluated based on two masks: one with all the 68 landmarks and one with 51 which are a subset of the 68 ones by removing the boundary (jaw) points. Figure 8 shows the evolution of the subspace for an indicative MMI video. The initial and final subspace are visualized in the top and bottom rows of the figure respectively. As can be seen, the initial bases display misaligned and blurred faces. However, in the resulting subspace, the facial areas are distinctive and clear. We think that this improvement is significant given the automatic nature of the proposed method and the fact that we use pixel intensities for the appearance representation and not any other sophisticated descriptor. Moreover, note that the convergence demonstrated by Figs. 7a,7b and 8 is achieved in only 5 global iterations of the method.

Finally, we conduct an experiment to compare the fitting accuracy of ARCA with three other landmark localization methods trained on manual annotations.



(a) (b) (c) (d)

Fig. 7. Face alignment results. 7a: Plot of the mean cost function error of Eq. 13 over all MMI videos per iteration. 7b: Plot of the mean normalized RMSE over all MMI videos per iteration. 7c, 7d: Comparison of the fitting accuracy of ARCA with methods trained on manual annotations for MMI and UNS respectively.

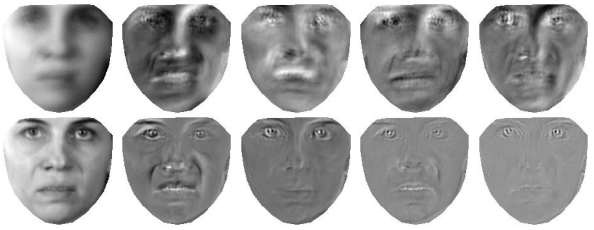


Fig. 8. Indicative example of the subspace evolution on an MMI video. *Top row:* Initial subspace. *Bottom row:* Final subspace after five iterations.

The first one is a person specific Active Appearance Model (AAM) trained using a small number of images for each subject. The second is a generic AAM trained on hundreds of "in-the-wild" images (captured in totally unconstrained conditions) from LFPW database [43]. The third methodology is Supervised Descent Method (SDM) [5], which uses the powerful SIFT features. For this technique, we utilize the implementation provided by the authors which has pretrained models built on thousands of images. We use the same initialization for all methods except SDM, for which we use the built-in initialization technique included in the online implementation. Figures 7c and 7d show the results on MMI and UNS databases respectively. ARCA performs better than the generic AAM. Moreover, it has worse performance than SDM and it is more robust but less accurate than the person-specific AAM. Note that the initialization of SDM is much better than the one of the rest of the methods, which partially explains the performance difference. We think that these results are remarkable given the automatic character of the proposed method and the fact that it is based on pixel intensities and not on any other powerful feature-based representation.

5 Conclusions

Contrary to what is practised in facial behaviour analysis, we show that it is possible to extract low-dimensional features that can capture the dynamics of the behaviour and jointly perform landmark localization. To do so we have introduced ARCA, Autoregressive component analysis, and we show that it possible to combine it with a motion model governt by a simple sparse shape model.

Acknowledgements. The work of Lazaros Zafeiriou has been funded by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 288235 (FROG). The work of Epameinondas Antonakos and Stefanos Zafeiriou was funded in part by the EPSRC project EP/J017787/1 (4DFAB). The work by Maja Pantic was funded in part by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA).

References

1. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 31(1), 39–58 (2009)
2. Zhou, F., De la Torre, F.: Canonical time warping for alignment of human behavior. In: Conference on Neural Information Processing Systems (NIPS), pp. 2286–2294 (2009)
3. Nicolaou, M.A., Pavlovic, V., Pantic, M.: Dynamic probabilistic cca for analysis of affective behaviour. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 98–111. Springer, Heidelberg (2012)
4. Tzimiropoulos, G., Alabort-i-Medina, J., Zafeiriou, S., Pantic, M.: Generic active appearance models revisited. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part III. LNCS, vol. 7726, pp. 650–663. Springer, Heidelberg (2013)
5. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Proceedings of Int'l Conf. on Computer Vision and Pattern Recognition (CVPR) (2013)
6. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: IEEE Proceedings of Int'l Conf. on Computer Vision and Pattern Recognition (CVPR) (2013)
7. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing (JIVC) 28(5), 807–813 (2010)
8. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: IEEE Proceedings of Int'l Conf. on Computer Vision and Pattern Recognition Workshop (CVPR-W 2013), 5th Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2013), Portland Oregon, USA (June 2013)
9. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: IEEE Proceedings of Int'l Conf. on Computer Vision Workshop (ICCV-W 2013), 300 Faces in-the-Wild Challenge (300-W), Sydney, Australia (December 2013)
10. Zhou, F., De la Torre, F., Cohn, J.F.: Unsupervised discovery of facial events. In: IEEE Proceedings of Int'l Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2574–2581. IEEE (2010)
11. Zhou, F., De la Torre, F., Hodgins, J.K.: Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 35(3), 582–596 (2013)
12. Antonakos, E., Pitsikalis, V., Rodomagoulakis, I., Maragos, P.: Unsupervised classification of extreme facial events using active appearance models tracking for sign language videos. In: IEEE Proceedings of Int'l Conf. on Image Processing (ICIP), Orlando, FL, USA (October 2012)
13. Zhang, W., Shan, S., Chen, X., Gao, W.: Local gabor binary patterns based on mutual information for face recognition. International Journal of Image and Graphics 7(04), 777–793 (2007)
14. Ha, S.W., Moon, Y.H.: Multiple object tracking using sift features and location matching. International Journal of Smart Home 5(4) (2011)
15. Zafeiriou, L., Nicolaou, M.A., Zafeiriou, S., Nikitidis, S., Pantic, M.: Learning slow features for behaviour analysis. In: IEEE Proceedings of Int'l Conf. on Computer Vision (ICCV) (November 2013)

16. Zhou, F., De la Torre, F.: Generalized time warping for multi-modal alignment of human motion. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
17. Rue, H., Held, L.: Gaussian Markov random fields: theory and applications. CRC Press (2004)
18. Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 34(11), 2233–2246 (2012)
19. Zhao, C., Cham, W.K., Wang, X.: Joint face alignment with a generic deformable face model. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 561–568. IEEE (2011)
20. Sagonas, C., Panagakis, Y., Zafeiriou, S., Pantic, M.: Raps: Robust and efficient automatic construction of person-specific deformable models. In: Proceedings of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR 2014) (June 2014)
21. De la Torre, F., Black, M.J.: Robust parameterized component analysis: theory and applications to 2d facial appearance models. Computer Vision and Image Understanding 91(1), 53–71 (2003)
22. De la Torre, F., Nguyen, M.H.: Parameterized kernel principal component analysis: Theory and applications to supervised and unsupervised image alignment. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
23. Cheng, X., Fookes, C., Sridharan, S., Saragih, J., Lucey, S.: Deformable face ensemble alignment with robust grouped-l1 anchors. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013), pp. 1–7 (2013)
24. Cheng, X., Sridharan, S., Saragih, J., Lucey, S.: Rank minimization across appearance and shape for aam ensemble fitting. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 577–584. IEEE (2013)
25. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Proceedings of Int'l Conf. on Computer Vision and Pattern Recognition (CVPR) (2001)
26. Lampert, C.H., Blaschko, M.B., Hofmann, T.: Efficient subwindow search: A branch and bound framework for object localization. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 31(12), 2129–2142 (2009)
27. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Proceedings of Int'l Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886 (2012)
28. Orozco, J., Martinez, B., Pantic, M.: Empirical analysis of cascade deformable models for multi-view face detection. In: IEEE Proceedings of Int'l Conf. on Image Processing (ICIP) (2013)
29. Jiang, T., Jurie, F., Schmid, C.: Learning shape prior models for object matching. In: IEEE Proceedings of Int'l Conf. on Computer Vision and Pattern Recognition (CVPR) (2009)
30. Kokkinos, I., Yuille, A.: Unsupervised learning of object deformation models. In: IEEE Proceedings of Int'l Conf. on Computer Vision (ICCV) (2007)
31. Yang, J., Frangi, A.F., Yang, J.Y., Zhang, D., Jin, Z.: Kpca plus lda: a complete kernel fisher discriminant framework for feature extraction and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 27(2), 230–244 (2005)

32. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. Journal of Computational and Graphical Statistics 15(2), 265–286 (2006)
33. Baker, S., Matthews, I.: Lucas-kanade 20 years on: A unifying framework. International Journal of Computer Vision (IJCV) 56(3), 221–255 (2004)
34. Jolliffe, I.: Principal component analysis. Wiley Online Library (2005)
35. Welling, M.: Fisher linear discriminant analysis. Department of Computer Science. University of Toronto 3 (2005)
36. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
37. He, X., Niyogi, P.: Locality preserving projections. In: NIPS, vol. 16, pp. 234–241 (2003)
38. Roweis, S., Ghahramani, Z.: A unifying review of linear gaussian models. Neural Computation 11(2), 305–345 (1999)
39. Wiskott, L., Sejnowski, T.J.: Slow feature analysis: Unsupervised learning of invariances. Neural Computation 14(4), 715–770 (2002)
40. Valstar, M.F., Pantic, M.: Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In: Proceedings of Int'l Conf. on Language Resources and Evaluation (LREC), Workshop on EMOTION, Malta (May 2010)
41. Valstar, M.F., Pantic, M.: Mmi facial expression database, `http://www.mmifacedb.com/`
42. Dibeklioglu, H., Salah, A.A., Gevers, T.: Uva-nemo smile database, `http://www.uva-nemo.org/`
43. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: IEEE Proceedings of Int'l Conf. on Computer Vision and Pattern Recognition (CVPR) (2011)