A Joint Discriminative Generative Model for Deformable Model Construction and Classification

Ioannis Marras[1], Symeon Nikitidis[2], Stefanos Zafeiriou[1] and Maja Pantic[1,3]
[1] Department of Computing, Imperial College London, UK
[2] Yoti Ltd, London, UK, e-mail: symeon.nikitidis@yoti.com
[3] Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands

***Abstract*— Discriminative classification models have been successfully applied for various computer vision tasks such as object and face detection and recognition. However, deformations can change objects coordinate space and perturb robust similarity measurement, which is the essence of all classification algorithms. The common approach to deal with deformations is either to seek for deformation invariant features or to develop models that describe objects deformations. However, the former approach requires a huge amount of data and a good amount of engineering to be properly trained, while the latter require considerable human effort in the form of carefully annotated data. In this paper, we propose a method that jointly learns with minimal human intervention a generative deformable model using only a simple shape model of the object and images automatically downloaded from the Internet, and also extracts features appropriate for classification. The proposed algorithm is applied on various classification problems such as "in-the-wild" face recognition, gender classification and eye glasses detection on data retrieved by querying into a web image search engine. We demonstrate that not only it outperforms other automatic methods by large margins, but also performs comparably with supervised methods trained on thousands of manually annotated data.**

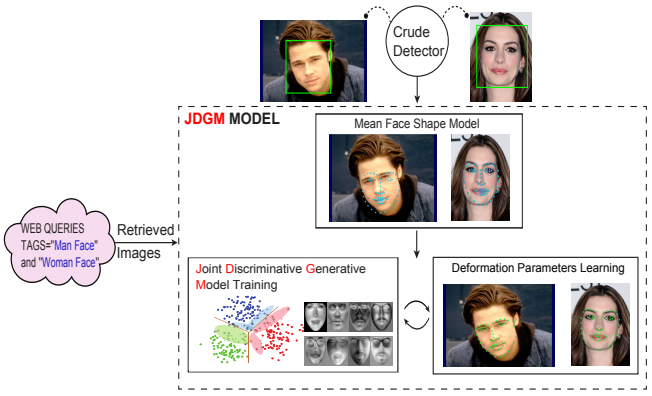

Fig. 1. Given a set of unconstrained facial images retrieved by performing the queries "Man Face" and "Woman Face" into a web image search engine, the proposed joint model using only the bounding boxes and a mean face shape model, iteratively aligns the retrieved images by leaning the deformation parameters, while extracts discriminant features and builds a classifier.

I. Introduction

Arguably, one of the problems that distinguishes computer vision discipline from machine learning is visual objects deformations modelling. Object deformations involve various shape and texture variations within an object class (e.g., the differences between parts of human faces), as well as, variations due to rigid or non-rigid movement of different object parts (e.g., head pose and facial expressions).

Objects deformations render the use of standard distance measures (such as the Euclidean distance), or kernels (such as Gaussian Radial Basis Function (GRBF) and polynomial kernels) not robust for measuring the similarity between the deformed data samples. Hence, on recognition problems involving visual data captured in unconstrained (also referred to as "in-the-wild") conditions, the direct application of popular distance based classifiers, such as Support Vector Machines (SVMs) [21], [19] and Relevance Vector Machines (RVMs) [1], results in a significantly degraded performance. Thus, an additional data pre-processing step is required to deal with objects deformations before the classification algorithm is applied.

Dealing with objects deformations has created a wealth of research which can be roughly divided into two categories: (1) methods that identify appropriate deformation invariant features of the objects and use these in order to train the classification algorithms and (2) methods that define statistical or physical deformable models of the visual objects at hand, in order to align the deformed objects prior to feature extraction and classification. In the first line of research fall methods that construct kernels invariant to certain rigid image transformations [14], as well as the family of deep learning and representation algorithms, such as Deep Convolutional Neural Networks (DCNN) [16], [12], [25] and Invariant Scattering Convolution Networks (ISCN) [5]. However, invariant kernels and ISCNs can only model simple objects transformations such as those caused by view angle rotation [14], while DCNNs in order to be robustly trained require huge amounts of data, and a good portion of engineering to handle the high computational complexity [25]. In addition, as it has been recently shown in [25], in challenging settings such as "in-the-wild" face recognition, DCNNs trained on millions of annotated data are unable to reach satisfactory performance and meticulously designed face alignment algorithms have to be applied to deal with face deformations. Finally, most of the above deformation

978-1-5090-4023-0/17/$31.00

invariant feature extraction algorithms are commonly combined with an SVM algorithm for classification.

In the second line of research fall methods that explicitly model objects deformations by defining a statistical deformable model. Notable examples include Active Appearance Models [7], Active Shape Models [24], Elastic Graph Matching [30] and Pictorial Structures [11], [32]. These models are applied in order to first align the data which are subsequently fed to the classifier. However, learning visual deformable models requires considerable human effort and still remains an expensive, tedious, labour intensive and prone to human errors procedure. More precisely, it requires both a careful selection of an image corpora that can efficiently exhibit the vast amount of objects variability and also the careful annotation of this corpora in terms of objects meaningful parts.

In this paper, we propose a methodology that is able to jointly learn with minimal supervision and annotation effort, classifiers and objects deformable models using data automatically collected from the Internet without requiring their detailed semantic annotation. Let us consider the following Internet image based application, where we want to train a facial gender classifier using the abundance of images available in the Internet which can be retrieved by applying a textual query (e.g., "man face" and "woman face") into a web image search engine such as Google and Bing, or photo sharing websites such as Flickr. Our application is further motivated by Figure 1. In this task we want to automatically and simultaneously learn (a) a facial deformable model associated with the image collection and (b) extract appropriate features and build a classifier for facial gender discrimination. If we merely use the obtained images to directly extract discriminant features and train a classifier, such as SVMs, the existence of facial deformations and misalignment, is guaranteed to lead to inaccurate results. To alleviate this problem we propose a joint methodology that has an interplay between a generative statistical model and a discriminative one, for facial image alignment and feature extraction/classification, on the latent space of the first. In particular, we aim to recover the deformation parameters (e.g., via the use of object shape parameters) in order to align the retrieved images using a generative model (i.e. Principal Component Analysis (PCA)) and at the same time to pursue a discriminative decomposition of the aligned data by coupling the objective of the generative model with a discriminative one (i.e. SVMs) that aims to separate data with maximum margin. This can be also interpreted as developing an SVM classifier with generative regularization terms and deformations parameters which enables the creation of the statistical deformable models.

The line of research that is closely related to the proposed work is that of image congealing [17] which refers to the problem of simultaneous aligning a set of images. However, the majority of the rigid and non-rigid image congealing algorithms use pure generative subspace models, such as PCA [9] or Robust PCA (RPCA) alternatives [4], [20], [23], [2]. On the other hand, we propose to the best of our knowledge the first methodology which simultaneously learns a subspace model that can be jointly used for deformable models construction, discriminant features extraction and classification [1].

Summarizing the novel contributions of the paper are the following:

- We propose a joint discriminative generative component analysis method which learns a subspace that not only best reconstructs the data but also provides low-dimensional features which can separate classes by maximum margin. Joint models are currently quite popular in computer vision. For instance, [18] proposes a joint model that learns a maximum margin classifier and extracts discriminant features in a single framework, while [11] jointly learns the statistical deformable model parameters and the parts locations in a discriminative manner. In addition, deep multi-task learning networks [31], [22] are among the state-of-the-art algorithms on various face analysis tasks.
- We show that is feasible to incorporate a deformation field in the proposed model, via a simple shape model, and, hence, to jointly align the set of images while learning the model parameters. The parameters of the model can be used to fit a deformable model on test images and at the same time perform classification. To the best of our knowledge such a model has not been presented before.
- We applied the proposed method on various recognition problems using unconstrained data, such as face recognition on LFW database [15] or facial gender classification and eye glass detection using data directly collected from the Internet using Google image search engine. We demonstrate that the proposed joint model outperforms the traditional ones that require the separate application of various algorithms for facial landmark localization, image alignment, feature extraction and classification.

II. Problem Formulation

In this section we present the proposed joint discriminative generative component analysis method that combines optimal data reconstruction with classes separation by maximum margin. We term this model *Joint Discriminative Generative Model (JDGM)* and develop an alternative optimization algorithm with close form solutions for its optimization. To facilitate our deformable model construction framework in the following we shall assume that our data are vectorized images. However, the presented component analysis algorithm can be applied on any kind of high dimensional data.

A. Joint Discriminative Generative Model

Given a set $\mathcal{X} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ of N training data pairs, where $\mathbf{x}_i \in \mathcal{R}^F, i = 1, ..., N$ are the F-dimensional input feature vectors each assigned a class label

[1]Our model can be also viewed as an application of the Occam's razor principle, which in simple terms requires to learn a model tailored to the task at hand and not a more general one, as those learnt by PCA or RPCA.

$y_i \in \{1, \ldots, K\}$ with K denoting the total number of classes, a multiclass SVM classifier [8] attempts to determine a set of K separating hyperplanes $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K\}$ where $\mathbf{w}_p \in \mathcal{R}^F, p = 1, ..., K$ is the normal vector of the p-th hyperplane that separates the training vectors of the p-th class from all the others with maximum margin by solving the following constraint optimization problem [8]:

$$\begin{aligned} \min_{\mathbf{w}_p, \xi_i} \quad & \frac{1}{2}\sum_{p=1}^{K} \mathbf{w}_p^T \mathbf{w}_p + C\sum_{i=1}^{N} \xi_i, \\ \text{s.t.:} \quad & \mathbf{w}_{y_i}^T \mathbf{x}_i - \mathbf{w}_p^T \mathbf{x}_i \geq b_i^p - \xi_i, \quad i = 1, \ldots, N. \end{aligned} \tag{1}$$

where $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^T$ are the slack variables, each one associated with a training sample, C is the term that penalizes the training error and $\mathbf{b}$ is the bias vector defined as $b_i^p = 1 - \delta_{y_i}^p$, where $\delta_{y_i}^p$ is the Kronecker delta function.

PCA is one of the most popular components analysis techniques that aims to find a set of orthonormal projection bases, stacked in the columns of matrix $\mathbf{U} \in \Re^{F \times M}$, such that the data reconstruction error is minimized or equivalently the variance of the projected samples is maximized:

$$\begin{aligned} \mathbf{U}_o &= \arg\min_{\mathbf{U}} \textstyle\sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{U}\mathbf{U}^T\mathbf{x}_i||_2^2 \\ &= \arg\min_{\mathbf{U}} ||\mathbf{X} - \mathbf{U}\mathbf{U}^T\mathbf{X}||_2^2 \\ &= \arg\max_{\mathbf{U}} ||\mathbf{U}^T\mathbf{X}||_2^2, \text{ s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}_M, \end{aligned} \tag{2}$$

where $\mathbf{X} \in \mathcal{R}^{F \times N}$ is the data matrix created by stacking samples $\mathbf{x}_i$ column-wise.

JDGM model aims to learn a subspace that best reconstructs the data, while providing appropriate for classification low dimensional features that can discriminate classes by maximum margin. Thus, we combine the above constrained optimization problems into a single objective function and derive the following minimax optimization problem:

$$\begin{aligned} \min_{\mathbf{w}_p, \xi_i} \max_{\mathbf{U}} \ \frac{1}{2}\sum_{p=1}^{K} \mathbf{w}_p^T \mathbf{w}_p \quad &+ \quad C\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} ||\mathbf{U}^T\mathbf{x}_i||_2^2 \\ \text{s.t.: } \mathbf{w}_{y_i}^T \mathbf{U}^T \mathbf{x}_i - \mathbf{w}_p^T \mathbf{U}^T \mathbf{x}_i \quad &\geq \quad b_i^p - \xi_i, \ i = 1, \ldots, N \\ \mathbf{U}^T\mathbf{U} \quad &= \quad \mathbf{I}_M \end{aligned}$$

where the separating hyperplane normal vector $\mathbf{w}_p \in \mathcal{R}^M$ is M-dimensional, since it separates samples in the subspace of $\mathbf{U}$, determined through the linear data projection $\acute{\mathbf{x}}_i = \mathbf{U}^T\mathbf{x}_i$ and $\mathbf{I}_M$ is an $M \times M$ identity matrix.

To solve the minimax optimization problem in (3) we consider an alternative optimization framework where we first compute the optimal decision hyperplanes for an initialized projection matrix $\mathbf{U}$ and subsequently, solve (3) for $\mathbf{U}$ while keeping the optimal normal vectors $\mathbf{w}_{p,o}$ fixed. Thus, to identify the optimal $\mathbf{w}_p$ we introduce positive Lagrange multipliers α_i^p and $\boldsymbol{\Lambda} \in \mathcal{R}^{M \times M} = [\Lambda_{i,j}]$ each associated with one inequality or orthonormality constraint, respectively and formulate the Lagrangian function $\mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\Lambda})$:

$$\begin{aligned} \mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}) &= \frac{1}{2}\sum_{p=1}^{K} \mathbf{w}_p^T \mathbf{w}_p + C\sum_{i=1}^{N} \xi_i \\ &+ \text{Tr}[\mathbf{X}^T\mathbf{U}\mathbf{U}^T\mathbf{X}] - \text{Tr}[\boldsymbol{\Lambda}(\mathbf{U}^T\mathbf{U} - \mathbf{I}_M] \\ &- \sum_{i=1}^{N}\sum_{p=1}^{K} \alpha_i^p \left[\left(\mathbf{w}_{y_i}^T - \mathbf{w}_p^T \right) \mathbf{U}^T \mathbf{x}_i + \xi_i - b_i^p \right], \end{aligned} \tag{3}$$

where Tr[.] is the matrix trace operator. To find the minimum over the primal variables $\mathbf{w}_p$ and $\boldsymbol{\xi}$ we require the partial derivatives of $\mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\Lambda})$ with respect to $\boldsymbol{\xi}$ and $\mathbf{w}_p$ to vanish, which yields the following equalities:

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \Rightarrow \sum_{p=1}^{k} \alpha_i^p = C, \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_p} = 0 \Rightarrow \mathbf{w}_{p,o} = \sum_{i=1}^{N} \left(\alpha_i^p - \sum_{p=1}^{K} \alpha_i^p \delta_{y_i}^p \right) \mathbf{U}^T \mathbf{x}_i. \tag{5}$$

Substituting terms from (4),(5) into (3), expressing the Lagrange multipliers in a vector form and performing the substitution $\mathbf{n}_i = C\mathbf{1}_{y_i} - \boldsymbol{\alpha}_i$, (where $\mathbf{1}_{y_i}$ is a K-dimensional vector with all its components equal to zero except of the y_i-th, which is equal to one) the minimax problem in (3) is equivalent to the following dual problem:

$$\begin{aligned} \min_{\mathbf{n}} \max_{\mathbf{U}} \quad & \frac{1}{2}\sum_{i,j}^{N} \mathbf{x}_i^T \mathbf{U}\mathbf{U}^T \mathbf{x}_j \mathbf{n}_i^T \mathbf{n}_j + \sum_{i=1}^{N} \mathbf{n}_i^T \mathbf{b}_i \\ & + \quad \text{Tr}[\mathbf{X}^T\mathbf{U}\mathbf{U}^T\mathbf{X}] - \text{Tr}[\boldsymbol{\Lambda}(\mathbf{U}^T\mathbf{U} - \mathbf{I}_M)], \\ \text{s.t.:} \quad \sum_{p=1}^{K} n_i^p &= 0, \quad n_i^p \leq \begin{cases} 0 & , \text{if } y_i \neq p \\ C & , \text{if } y_i = p \end{cases} \\ \forall \quad i &= 1, \ldots, N \quad , \quad p = 1, \ldots, K. \end{aligned} \tag{6}$$

Solving the above quadratic programming problem with the linear kernel function
$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{U}\mathbf{U}^T \mathbf{x}_j$ for $\mathbf{n}$, we subsequently obtain the separating hyperplanes from (5).

To optimize for $\mathbf{U}$ we remove from (6) term $\sum_{i=1}^{N} \mathbf{n}_i^T \mathbf{b}_i$, since it is independent of the optimized variable and solve the resulting trace optimization problem:

$$\begin{aligned} \max_{\mathbf{U}} \ & \text{Tr}[\mathbf{U}^T \sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T \mathbf{U}] + \text{Tr}[\mathbf{X}^T\mathbf{U}\mathbf{U}^T\mathbf{X}] \\ & - \text{Tr}[\boldsymbol{\Lambda}(\mathbf{U}^T\mathbf{U} - \mathbf{I}_M)]. \end{aligned} \tag{7}$$

Computing and setting the derivative with respect to $\mathbf{U}$ equal to zero we derive the following generalized eigenvalue problem:

$$\Big(\sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T + \mathbf{X}\mathbf{X}^T \Big) \mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}. \tag{8}$$

Thus, the projection bases of $\mathbf{U}$ correspond to the $(N-1)$ eigenvectors of matrix $\sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T + \mathbf{X}\mathbf{X}^T$ associated with the largest eigenvalues. Matrix $\sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T$ has a form similar to the Linear Discriminant Analysis (LDA)

between class covariance matrix, since it can be written as $\sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T = \mathbf{X}\mathbf{L}\mathbf{L}^T\mathbf{X}^T = \mathbf{A}\mathbf{A}^T$, where $\mathbf{A} \in \mathcal{R}^{F\times K}$ and $\mathbf{L} = [\mathbf{n}_1, \ldots, \mathbf{n}_K]^T \in \mathbf{R}^{N\times K}$ is created by stacking the vectors of the optimal Lagrange multipliers for each training sample row-wise. Since non-zero Lagrange multipliers correspond to support vectors which are the samples that reside closest to the decision boundary, matrix $\sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T$ is a robust estimator of the between class scatter evaluated using only the support vectors weighted by their associated Lagrange multiplier value. On the other hand, matrix $\mathbf{S}_t = \mathbf{X}\mathbf{X}^T$ is the data covariance matrix, thus their combination enables the interplay between extracting simultaneously generative and discriminant basis vectors.

III. JDGM with Generalized Orthogonality Constraints

On deformable models construction it is usually advantageous to whiten or "sphere" the data. In our model this can be directly incorporated in the optimization problem by the additional constraint $\mathbf{u}_k^T \mathbf{S}_t \mathbf{u}_k = 1$, where $\mathbf{u}_k$ is the k-th projection base of matrix $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M]$. Thus, to evaluate the data transformation matrix $\mathbf{U}$, while incorporating the additional orthogonality constraint, we develop an optimization algorithm motivated by [6] where we sequentially derive each projection base $\mathbf{u}_k$, $k = 1, \ldots, M$ so that it improves the objective function and at the same time is orthogonal to the $(k-1)$ previously derived bases. More precisely, to evaluate the k-th basis vector we consider the following minimax optimization problem:

$$\begin{aligned} \min_{\mathbf{w}_p, \xi_i} \max_{\mathbf{u}_k} \quad & \frac{1}{2}\sum_{p=1}^{K} \mathbf{w}_p^T \mathbf{w}_p + C\sum_{i=1}^{N} \xi_i, \\ \text{s.t.:} \quad & \mathbf{w}_{y_i}^T \mathbf{u}_k^T \mathbf{x}_i - \mathbf{w}_p^T \mathbf{u}_k^T \mathbf{x}_i \geq b_i^p - \xi_i, \\ & \mathbf{u}_k^T \mathbf{u}_1 = \mathbf{u}_k^T \mathbf{u}_2 = \cdots = \mathbf{u}_k^T \mathbf{u}_{k-1} = 0, \\ & \mathbf{u}_k^T \mathbf{S}_t \mathbf{u}_k = 1. \end{aligned} \tag{9}$$

Deriving the dual formulation of (9) with respect to the primal variables $\mathbf{w}_p$ and ξ_i as in (6) and setting $\mathbf{S}_b = \sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T$, we obtain:

$$\begin{aligned} \min_{\mathbf{n}} \max_{\mathbf{u}_k} \quad & \frac{1}{2}\mathbf{u}_k^T \mathbf{S}_b \mathbf{u}_k + \sum_{i=1}^{N} \mathbf{n}_i^T \mathbf{b}_i \\ \text{s.t.:} \quad & \sum_{p=1}^{K} n_i^p = 0, \quad n_i^p \leq \begin{cases} 0 & , \text{if } y_i \neq p \\ C & , \text{if } y_i = p \end{cases} \\ & \forall \quad i = 1, \ldots, N \quad , \quad p = 1, \ldots, K. \\ & \mathbf{u}_k^T \mathbf{u}_1 = \mathbf{u}_k^T \mathbf{u}_2 = \cdots = \mathbf{u}_k^T \mathbf{u}_{k-1} = 0, \\ & \mathbf{u}_k^T \mathbf{S}_t \mathbf{u}_k = 1. \end{aligned} \tag{10}$$

The solution for the optimal decision hyperplanes is computed by fixing the transformation bases and solving the resulting QP problem for $\mathbf{n}$ as in (6). To compute the k-th orthogonal basis $\mathbf{u}_k$ we fix the optimal normal vectors $\mathbf{w}_{p,o}$ and consider the maximization problem:

$$\begin{aligned} \mathbf{u}_k = \arg\max_{\mathbf{u}} \quad & \frac{1}{2}\mathbf{u}_k^T \mathbf{S}_b \mathbf{u}_k + \sum_{i=1}^{N} \mathbf{n}_i^T \mathbf{b}_i \\ \text{s.t.: } \mathbf{u}_k^T \mathbf{u}_1 = \mathbf{u}_k^T \mathbf{u}_2 = \cdots = \mathbf{u}_k^T \mathbf{u}_{k-1} &= 0, \\ \mathbf{u}_k^T \mathbf{S}_t \mathbf{u}_k &= 1. \end{aligned} \tag{11}$$

Incorporating the constraints we develop the following Lagrangian function:

$$\begin{aligned} \mathcal{L}(\mathbf{u}_k, \lambda, \boldsymbol{\mu}) = & \frac{1}{2}\mathbf{u}_k^T \mathbf{S}_b \mathbf{u}_k - \lambda(\mathbf{u}_k^T \mathbf{S}_t \mathbf{u}_k - 1) \\ & - \mu_1 \mathbf{u}_k^T \mathbf{u}_1 - \mu_2 \mathbf{u}_k^T \mathbf{u}_2 - \cdots - \mu_{k-1} \mathbf{u}_k^T \mathbf{u}_{k-1}, \end{aligned} \tag{12}$$

where computing and setting its derivative with respect to $\mathbf{u}_k$ equal to zero we derive the following equality:

$$\mathbf{S}_b \mathbf{u}_k - 2\lambda \mathbf{S}_t \mathbf{u}_k - \mu_1 \mathbf{u}_1 - \mu_2 \mathbf{u}_2 - \cdots - \mu_{k-1} \mathbf{u}_{k-1} = 0. \tag{13}$$

Multiplying (13) from the left hand side by $\mathbf{u}_k^T$ and solving for λ we derive:

$$\lambda = \frac{\mathbf{u}_k^T \mathbf{S}_b \mathbf{u}_k}{2\mathbf{u}_k^T \mathbf{S}_t \mathbf{u}_k}, \tag{14}$$

while multiplying the left hand side of (13) by $\mathbf{u}_1 \mathbf{S}_t^{-1}$, $\mathbf{u}_2 \mathbf{S}_t^{-1}, \ldots, \mathbf{u}_{k-1} \mathbf{S}_t^{-1}$, summing up the resulting equations and setting $\mathbf{U}^{(k-1)} = [\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}]$, $\boldsymbol{\mu}^{(k-1)} = [\mu_1, \ldots, \mu_{k-1}]$ and $\boldsymbol{\Theta}^{(k-1)} = [\mathbf{U}^{(k-1)}]^T \mathbf{S}_t^{-1} \mathbf{U}^{(k-1)}$ we derive:

$$\boldsymbol{\mu}^{(k-1)} = [\boldsymbol{\Theta}^{(k-1)}]^{-1} [\mathbf{U}^{(k-1)}]^T \mathbf{S}_t^{-1} \mathbf{S}_b \mathbf{u}_k. \tag{15}$$

Finally, multiplying the left hand side of (13) by $\mathbf{S}_t^{-1}$ and substituting terms according to (15) we derive the generalized eigenvalue problem:

$$\frac{1}{2}\left(\mathbf{I}_F - \mathbf{S}_t^{-1} \mathbf{U}^{(k-1)} [\boldsymbol{\Theta}^{(k-1)}]^{-1} [\mathbf{U}^{(k-1)}]^T\right) \mathbf{S}_t^{-1} \mathbf{S}_b \mathbf{u}_k = \lambda \mathbf{u}_k \tag{16}$$

We compute $\mathbf{u}_1$ as the eigenvector of matrix $\mathbf{S}_t^{-1}\mathbf{S}_b$ associated with the largest eigenvalue and evaluate sequentially the remaining $(N-1)$ orthogonal bases $\mathbf{u}_k$ as the eigenvectors with the largest eigenvalues of $\frac{1}{2}\left(\mathbf{I}_F - \mathbf{S}_t^{-1} \mathbf{U}^{(k-1)} [\boldsymbol{\Theta}^{(k-1)}]^{-1} [\mathbf{U}^{(k-1)}]^T\right) \mathbf{S}_t^{-1} \mathbf{S}_b$.

IV. Deformation Parameters Learning and Image Alignment

The proposed framework is a general method for automatic deformable model construction and thus can be applied on any deformable object. Here we explicitly focus on facial images and on building facial deformable models, since various annotated in-the-wild facial databases, shape models and algorithms for facial landmark localization are available which can facilitate our quantitative evaluation.

To automatically learn the shape deformation parameters for image alignment we only use a shape prior and the initialized bounding boxes. More formally, given a set of N training images $\{\mathbf{I}^i\}$, $i = 1, \ldots, N$ and a statistical shape model $\{\bar{\mathbf{s}}, \mathbf{B}\}$ where $\bar{\mathbf{s}} = \{(x_1, y_1), \ldots, (x_L, y_L)\}$ is the mean shape model and (x_i, y_i) are the coordinates of the i-th landmark point (we consider an $L = 68$ landmark points

facial shape model) and matrix $\mathbf{B} \in \Re^{2L\times(4+N_s)}$ stores the shape eigenvectors where the first four correspond to the global similarity transform, while the rest are computed performing PCA on an available set of facial shapes.

A new shape instance is generated as a perturbation of the mean shape $\bar{\mathbf{s}}$ by linearly combining the eigenvectors in $\mathbf{B}$ weighted by the parameters $\mathbf{p} = [p_1, \ldots, p_{4+N_S}]^T$ as $\mathbf{s}_p = \bar{\mathbf{s}} + \mathbf{B}\mathbf{p}$. Moreover, we define a motion model as the warp function $W(\mathbf{x}, \mathbf{p})$, (which for simplicity we denote as $W(\mathbf{p})$) that maps each point within the mean shape ($\mathbf{x} \in \bar{\mathbf{s}}$) to its corresponding location in a shape instance. Thus, each image warping to the mean shape given a shape estimate of the displayed face ($\{\mathbf{s}_i\},\ i = 1, \ldots, N$), returns N appearance vectors $\{\mathbf{I}^i(W(\mathbf{p}_i))\},\ i = 1, \ldots, N$ of size $F\times 1$, where F is the total number of pixels that lie inside the mean shape. We also denote as $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N]$ the $(4+N_s)\times N$ matrix containing the shape parameters for each image. To estimate the motion parameters $\mathbf{p}_i,\ i = 1, \ldots, N$ we consider the following optimization problem where we aim to minimize the ℓ_2^2 norm fitting error over all images using the generative discriminative bases $\mathbf{U}$ derived by the JDGM model:

$$\min_{\{\mathbf{p}_i,\ i=1,\ldots,N\}} \sum_{i=1}^{N} ||\mathbf{x}_i(W(\mathbf{p}_i)) - (\mathbf{U}\mathbf{c}_i + \boldsymbol{\mu})||_2^2 \qquad (17)$$

where $\mathbf{c}_i \in \Re^{M\times 1},\ i = 1, \ldots, N$ are the linear combination weights and $\boldsymbol{\mu} \in \Re^{F\times 1}$ is the data mean vector. The optimization problem in (17) is solved in an alternative optimization manner as in [27], with $\mathbf{c}_i = \mathbf{U}^T(\mathbf{x}_i(W(\mathbf{p}_i)) - \boldsymbol{\mu})$, while the update of $\mathbf{p}_i$ is performed as in [26].

The proposed algorithm optimizes the proposed cost function using a standard block coordinate optimization approach. Even though it is not feasible to provide a mathematical proof of convergence, we observed empirical convergence in all cases. In all experiments, JDGM converged after a few tens of iterations (around 30) without facing any convergence issues.

V. Experimental Results

In the experimental comparisons we demonstrate that the proposed joint model is able to (a) correctly locate the facial landmarks from a crude face detector so as to perform image alignment (b) extract appropriate features for classification by performing a discriminative decomposition of the aligned data and (c) build a maximum margin classifier on the identified latent space. We compare the automatic JDGM model against the current standard approach on recognition applications involving in-the-wild facial images where various algorithms for facial landmarks localization, image alignment, feature extraction and finally classification are independently treated and considered as different modules in the pipeline of the general recognition application. To do so, we consider three different recognition problems on unconstrained facial data namely, face recognition on LFW dataset and eye glasses detection and facial gender classification on facial images collected from the Internet by performing textual queries in Flickr and Google web image search engine. The detailed descriptions of each dataset are given in subsections V-A, V-B and V-C.

In our experiments the facial bounding boxes of all images were derived by applying the Viola-Jones object detection algorithm [29], while we employed a shape model trained on 50 shapes of Multi-Pie database [13]. This shape model contains $N_s = 10$ eigenvectors and the mean shape has a resolution of 86×87 pixels, thus the initial dimensionality of all our data is $F = 7,482$.

We evaluate JDGM ability to automatically build a face shape model (i.e. facial landmark localization accuracy) and its feature extraction and classification performance, comparing against three different algorithms. More precisely, for the baseline algorithm in our comparison we directly exploited the detected facial regions (without aligning the images) which were anisotropically scaled to a fixed size of 86×87 pixels, performed PCA and LDA for dimensionality reduction and the extracted features were subsequently fed into a linear SVM algorithm [10] for classification. A purely automatic algorithm that we also considered in our comparison is the state-of-the-art image congealing *"TIP-CA"* algorithm in [9] that combines PCA with AAM fitting which we appropriately modified so as to learn the shape deformation parameters instead of the image affine transform. On the aligned facial images derived by TIP-CA we used only the pixels inside the resulting reference shapes, performed PCA and LDA for dimensionality reduction and used the same linear SVM algorithm for classification. For the third algorithm in the comparison which we term *"DRMF+SVM"*, we aligned the facial images exploiting the facial landmarks localizations derived by applying the state-of-the-art Robust Discriminative Response Map Fitting (DRMF) algorithm [3], trained on thousands of manually annotated facial images and similarly used only the pixels inside the shapes to perform PCA and LDA and fed the low dimensional discriminant representations to a linear SVM algorithm for classification. Finally, except of the purely automatic JDGM model where deformable shape fitting is initialized using the mean shape, we also initialized our model using the landmark localizations derived by DRMF developing an algorithm which we term *"JDGM+DRMF"*. Moreover, in all experiments we considered the image intensity values and the Image Gradient Orientations (IGOs) for robust subspace analysis as demonstrated in [28].

A. Face Recognition on LFW Dataset

LFW image dataset realistically simulates the variability evident to in-the-wild face recognition problems and is the standard benchmark on face verification. In total LFW consists of $13,233$ facial images depicting $5,749$ different individuals of which $4,069$ have a single image in the database. Thus, in order to perform face recognition, we employed a subset of the available data considering only the face images of those individuals that have more than 30 samples, sufficient to contribute both to the training and test sets. The dataset we considered consists in total of $2,310$ face images from 32 individuals where for each subject a

different number of samples were available varying from 30 to 530 images. To form our testing set we randomly selected half of the images available for each individual for training and the remaining for testing.

To measure the landmark localization accuracy we compute the point-to-point RMSE normalized with respect to the face size. Thus, the RMSE between the fitted shape $\mathbf{s}^f$ and the groundtruth shape $\mathbf{s}^g$ is evaluated as: $RMSE = \frac{\sum_{i=1}^{N}\sqrt{(x_i^f - x_i^g)^2 + (y_i^f - y_i^g)^2}}{N_s * d}$, where $d = (\max_x \mathbf{s}^g - \min_x \mathbf{s}^g + \max_y \mathbf{s}^g - \min_y \mathbf{s}^g)$. Figure 2 shows the normalized RMSE fitting curves for both appearance representation features where the proposed automatic joint model is compared against the TIP-CA and the state-of-the-art DRMF algorithm which however, is trained on manually annotated data. As can be seen, the proposed automatic JDGM method outperforms TIP-CA method, while it achieves comparable performance with DRMF algorithm. It is also important to highlight that JDGM+DRMF algorithm which exploits the landmark localizations derived by DRMF in order to initialize our deformable shape fitting process, achieved the best performance as it was able to further improve landmarks localization in images where the provided fittings were not accurate.

Table I summarizes the highest face recognition rates achieved by each method in the comparison. The proposed JDGM automatic model outperforms all other automatic algorithms by large margins, while the highest performance was achieved by JDGM+DRMF algorithm which also obtained the best fitting accuracy in the database.

TABLE I

FACE RECOGNITION ACCURACY RATES IN LFW IMAGE DATASET.

	Automatic Methods			Methods Requiring Landmarks Annotation	
Methods	Baseline	TIP-CA	JDGM	DRMF+SVM	JDGM+DRMF
Intensities	71.8%	74.6%	**86.9%**	88.5%	**90.4%**
IGOs	79.9%	84.1%	**91.1%**	93.3%	**95.8%**

B. Eye Glasses Detection

Next we consider a recognition problem using unconstrained images automatically collected from the Internet, where we aim to recognize whether the depicted subjects wear glasses or not. To do so, we created an image dataset depicting 20 different celebrities who usually appear wearing glasses (we considered both eye and sun glasses) and aim to automatically build a deformable face shape model for image alignment and also build a classifier and extract discriminant features appropriate for the eye glasses classification problem. To create our database we performed textual queries such as *"Samuel Jackson wearing glasses"* or *"Samuel Jackson without glasses"* in Google images and collected for each celerity 50 retrieved images containing approximately equal number of samples for each class. Thus, our celebrities image collection consists in total of $1,000$ images which were randomly split into half to form our training and test sets. Figure 3 shows some of the collected images where the facial landmarks have been localized using the proposed automatic JDGM model.

Fig. 3. Sample images of the collected celebrities dataset for the eye glasses detection problem. Facial landmarks have been localized by the proposed automatic JDGM model.

To further investigate the effect of the joint generative discriminative component analysis model in AAM fitting we computed the normalized RMSE considering only the facial landmarks around the eyes and the eyebrows which are the facial regions discriminant for the eye glasses classification problem. Figure 4 plots the fitting curves for both appearance features where as can be seen the proposed method achieved smaller fittings error. It is important to highlight that by directly comparing the fitting errors derived by TIP-CA algorithm that combines PCA with AAM fitting and the proposed JDGM model we verify that the discriminant information incorporated in the learnt subspace by JDGM significantly assists Gauss-Newton optimization algorithm to reach a better local minimum. Table II also summarizes the obtained classification results. JDGM+DRMF algorithm achieved the best recognition performance, while the purely automatic JDGM model also attained competitive performance outperforming TIP-CA algorithm by large margins.

TABLE II

EYE GLASSES DETECTION ACCURACY RATES IN THE COLLECTED CELEBRITIES DATABASE.

	Automatic Methods			Methods Requiring Landmarks Annotation	
Methods	Baseline	TIP-CA	JDGM	DRMF+SVM	JDGM+DRMF
Intensities	65.2%	69.3%	**82.6%**	84.8%	**85.2%**
IGOs	73.3%	77.0%	**89.6%**	92.2%	**93.7%**

C. Facial Gender Classification in-the-wild

For our final set of experiments we considered the problem of facial gender classification also in a set of in-the-wild facial images collected from the Internet. In order to form our image collection we used Flickr image sharing website where we applied the textual queries *"Man face"* and *"Woman face"* and collected 500 of the retrieved for each query images. Thus, in total our dataset comprises of $1,000$ facial images which were randomly split into half to form the training and test sets.

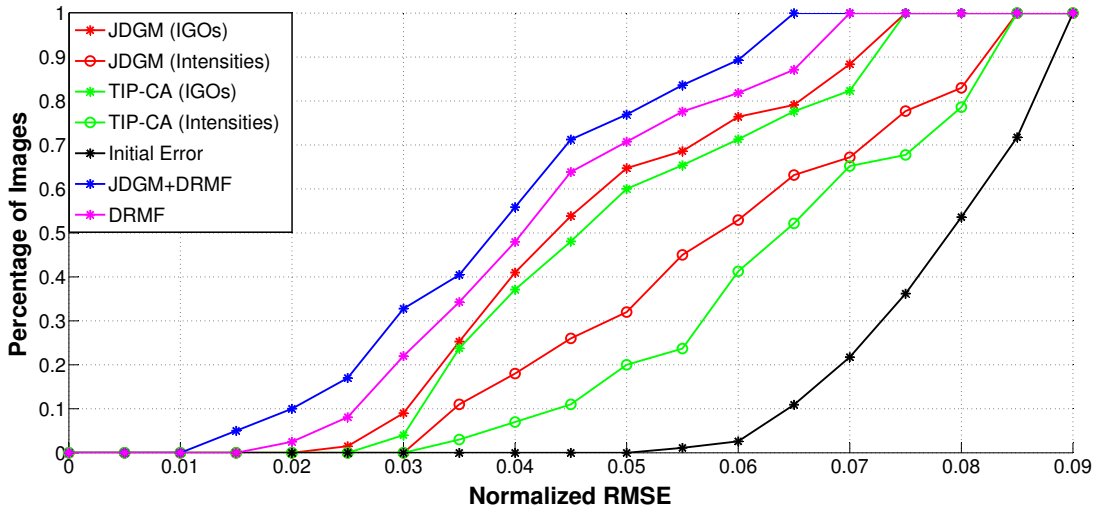

Fig. 2. Comparison of the normalized RMSE fitting curves in LFW image dataset.

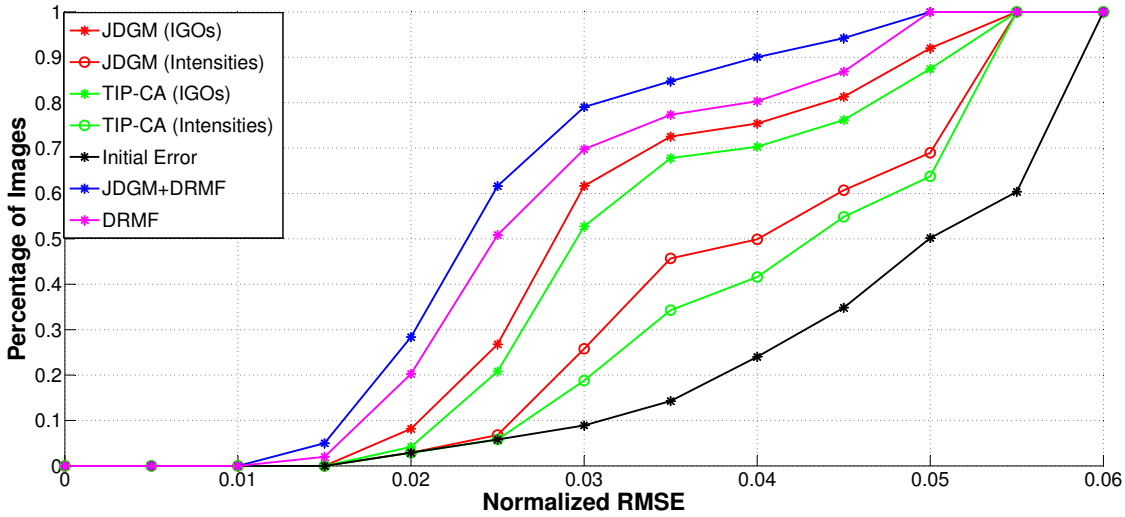

Fig. 4. Comparison of the normalized RMSE fitting curves in the collected celebrities dataset considering only the facial landmarks that correspond in the discriminant facial areas around the eyes and the eyebrows.

Table III summarizes the obtained classification results which are consistent with those obtained on the other datasets. JDGM model not only outperformed the other automatic methods but also attained competitive performance against models trained on manual annotations which we think is remarkable.

TABLE III
FACIAL GENDER CLASSIFICATION ACCURACY IN THE COLLECTED DATASET.

	Automatic Methods			Methods Requiring Landmarks Annotation	
Methods	Baseline	TIP-CA	JDGM	DRMF+SVM	JDGM+DRMF
Intensities	68.5%	71.1%	**88.4%**	89.6%	**90.7%**
IGOs	72.6%	77.4%	**89.1%**	91.5%	**93.9%**

VI. CONCLUSION

The current standard approach on recognition applications involving in-the-wild images of deformable objects requires the combination of various algorithms for landmarks localization, image alignment, feature extraction and classification which however, are independently treated and considered as different modules in the pipeline of the general recognition application. Contrary to that, we propose a model that jointly learns a generative deformable model using only a simple shape model of the object and a crude detector, and also extracts features appropriate for classification. Experiments on unconstrained facial data, some of which were directly collected from the Internet by performing textual queries in web image search engines, demonstrated that the proposed automatic model not only outperforms other automatic models by large margins but also was able to achieve comparable performance with models trained on thousands of manually annotated data.

VII. Acknowledgements

The work of S. Zafeiriou and M. Pantic was partially funded by EPSRC project FACER2VM (EP/N007743/1). The work of S. Nikitidis and I. Marras was funded by EPSRC 4D Facial Behaviour Analysis for Security (4DFAB, EP/J017787/1).

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *CVPR*, volume 2, pages 882–888. IEEE, 2004.

[2] E. Antonakos and S. Zafeiriou. Automatic construction of deformable models in-the-wild. In *CVPR*, pages 1813–1820. IEEE, 2014.

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, pages 3444–3451. IEEE, 2013.

[4] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *IEEE TPAMI*, 26(10):1380–1384, 2004.

[5] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE TPAMI*, 35(8):1872–1886, 2013.

[6] D. Cai, X. He, J. Han, and H.-J. Zhang. Orthogonal laplacianfaces for face recognition. *IEEE TIP*, 15(11):3608–3614, 2006.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE TPAMI*, 23(6):681–685, 2001.

[8] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2001.

[9] W. Deng, J. Hu, J. Lu, and J. Guo. Transform-invariant pca: A unified approach to fully automatic face alignment, representation, and recognition. *TPAMI*, 36(6):1275–1284, 2013.

[10] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587. IEEE, 2014.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing (JIVC)*, 28(5):807–813, 2010.

[14] O. C. Hamsici and A. M. Martinez. Rotation invariant kernels and their application to shape analysis. *IEEE TPAMI*, 31(11):1985–1999, 2009.

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[17] E. G. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE TPAMI*, 28(2):236–250, 2006.

[18] S. Nikitidis, S. Zafeiriou, and M. Pantic. Merging svms with linear discriminant analysis: a combined model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1067–1074, 2014.

[19] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *CVPR*, pages 130–136. IEEE, 1997.

[20] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE TPAMI*, 34(11):2233–2246, 2012.

[21] M. Pontil and A. Verri. Support vector machines for 3d object recognition. *IEEE TPAMI*, 20(6):637–646, 1998.

[22] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv:1603.01249*, 2016.

[23] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Raps: Robust and efficient automatic construction of person-specific deformable models. In *CVPR*, pages 1789–1796. IEEE, 2014.

[24] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, 2011.

[25] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708. IEEE, 2014.

[26] G. Tzimiropoulos, J. Alabort-i Medina, S. Zafeiriou, and M. Pantic. Generic active appearance models revisited. In *ACCV*, pages 650–663. Springer, 2013.

[27] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *ICCV*, pages 593–600. IEEE, 2013.

[28] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Subspace learning from image gradient orientations. *IEEE TPAMI*, 34(12):2454–2466, 2012.

[29] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages 511–518. IEEE, 2001.

[30] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. Von Der Malsburg. Face recognition by elastic bunch graph matching. *IEEE TPAMI*, 19(7):775–779, 1997.

[31] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.

[32] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886. IEEE, 2012.