Joint Head Pose / Soft Label Estimation for Human Recognition *In-The-Wild*

Hugo Proença, *Senior Member, IEEE*, João C. Neves, *Student Member, IEEE*, Silvio Barra, Tiago Marques and Juan C. Moreno

***Abstract*—Soft biometrics have been emerging to complement other traits and are particularly useful for poor quality data. In this paper, we propose an efficient algorithm to estimate human head poses and to infer soft biometric labels based on the 3D morphology of the human head. Starting by considering a set of pose hypotheses, we use a learning set of head shapes synthesized from anthropometric surveys to derive a set of 3D *head centroids* that constitutes a metric space. Next, representing queries by sets of 2D head landmarks, we use projective geometry techniques to rank efficiently the joint 3D head centroids / pose hypotheses according to their likelihood of matching each query. The rationale is that the most likely hypotheses are *sufficiently* close to the query, so a good solution can be found by convex energy minimization techniques. Once a solution has been found, the 3D head centroid and the query are assumed to have similar morphology, yielding the soft label. Our experiments point toward the usefulness of the proposed solution, which can improve the effectiveness of face recognizers and can also be used as a privacy-preserving solution for biometric recognition in public environments.**

***Index Terms*—Soft Biometrics, Visual Surveillance, Homeland Security, Privacy-preserving Recognition.**

I. Introduction

IN biometrics research, one of the most challenging goals is the development of recognition systems that work in unconstrained (outdoor) scenarios and do not assume the subjects' willingness to be recognized. In such conditions, the acquired data has poor quality, with faces partially occluded, blurred, or misaligned (Fig. 1).

Fig. 1. Examples of images acquired by a visual surveillance system, composed by a wide-view camera feeding a pan-tilt-zoom device that collects data from moving and at-a-distance targets (up to 40 meters away).

The idea behind soft biometrics is to obtain "*characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals*" [16]. These characteristics not only complement strong biometric traits, but they also prune the set of identities for a query. Soft biometrics can also be regarded as a response to privacy / ethical issues in using biometrics in public places: it makes it possible to ignore the large majority of the identities in the scene and attempt positive recognition (e.g., with a face recognizer) only for the subjects with soft labels similar to the identities on a watch-list.

This paper describes an algorithm to infer jointly human head poses and soft labels in an efficient way based on poor-quality data. During the learning phase, anthropometric head surveys feed a stochastic process that generates a set of synthetic 3D head meshes representing the major features of a population. Such elements are the input of a self-organizing map that obtains a discretized representation of the feature space, i.e., a matrix of *centroid* heads with a key property; it preserves the topological properties of the input space and enables us to define the closeness of its elements (i.e., the similarity of head shapes). Considering the wildness of the data, we also generate a set of pose hypotheses. Next, all combinations of joint poses / head shape hypotheses are grouped and indexed using as a criterion the proximity of their projected head landmarks.

In classification, having a query represented by a set of head image landmarks (detected as described in [18] or [8]), we rank the set of hypotheses in approximate logarithmic time according to the similarity between the query and the joint pose / head shape 2D projections. The idea is that the most likely hypothesis is *sufficiently* close to the solution so that only slight changes in its parameterization are required to match the query faithfully. This way, local minima are neglected and convex optimization techniques are used to reach acceptable solutions. A convergence test determines whether the process stops or the next hypothesis is considered. The method described here uses some insights from [37] and [30], namely in the generation of the set of hypotheses and in using projective geometry techniques to evaluate them.

The remainder of this paper is organized as follows: Section II summarizes the related work. Sections III and IV give a detailed description of the learning and classification phases of the proposed algorithm. Section V describes the experiments carried out and discusses the corresponding results. Finally, Section VI concludes the paper.

All authors but Silvio Barra are with the IT: Instituto de Telecomunicações, Department of Computer Science, University of Beira Interior, Covilhã, Portugal, E-mail: {hugomcp, jcneves, tmarques, jcmb}@di.ubi.pt. Silvio Barra is with the Università degli Studi di Cagliari, Italy, E-mail: silvio.barra@unica.it.
This work was supported by FCT project UID/EEA/50008/2013.
Manuscript received April 19, 2005; revised December 27, 2012.

II. Related Work

A. Soft Biometrics

According to [40], soft biometric traits are classified into three families: 1) global traits, which regard demographic information (e.g., age, gender, and ethnicity); 2) body traits, which are concerned with the subjectÕs somatotype, i.e., their overall appearance (height or body volume); and 3) head traits, which analyze the regions that humans instinctively use to identify others (e.g., hair or eye color, nose or neck thickness, and ear shape / size).

Regarding global traits, Heckathorn *et* al. [11] measured lengths of wrists and forearms. Using the concept of *interchangeability of indicators*, they argued that combining multiple low accuracy measurements yields a highly accurate indicator. Jain and Park [17] used demographic information (gender and ethnicity) and facial marks (scars, moles and freckles) to improve face image matching and retrieval performance. An extended version of this work can be found in [32].

In terms of body traits, Lucas and Henneberg [23] concluded that, upon the availability of accurate anthropometric measurements, the body is actually more distinctive than the face when distinguishing humans. Previously, other works (e.g., Rice *et* al. [36]) concluded that identification based on body measurements can be as accurate as using the face. Moustakas *et* al. [29] suggested a framework based on height and stride length information to increase the effectiveness of a gait recognition system, integrating soft labels directly in the estimation of the matching score instead of the traditionally used score-level fusion. Drosou *et* al. [7] proposed a probabilistic framework for improving the recognition performance via soft labels (global and body-based), modelling the systematic intrinsic error of each measurement (e.g., due to clothing).

Finally, most works in the head traits family analyze the discriminability of hair / facial hair styles and lengths. Dass *et* al. [6] pre-aligned the images based on the position of the eyes and, using agglomerative clustering techniques, defined five groups of hairstyles according to hair density in image patches. Hewig *et* al. [13] observed that the typical hair styles are heavily correlated with global traits (gender and age), which might also be useful for identification.

A noteworthy conclusion was drawn by Reid *et* al. [35]: *comparative* descriptors (relative magnitude between subjects' measurements) have more discriminatory power than the absolute values themselves, and are particularly advantageous in terms of stability. Detailed information about soft biometrics can be found in two comprehensive surveys by Kim *et* al. [25] and Reid *et* al. [34].

B. Head Pose Estimation

The existing methods for head pose estimation can be divided into two main groups: 1) generative, by fitting parametric models to the query; and 2) discriminative, which are model-free and search for correspondences between image features and known pose configurations.

Generative models consider prior information about human kinematics and anthropometry to reduce the number of plausible configurations for a query. In this family of approaches, appearance template methods (e.g., fed by Gabor descriptors [38]), flexible models based on the elastic graph matching (e.g., [27]) or active appearance models (e.g., [42]) can be highlighted. Model fitting methods, based on generic 3D face [1] and ellipsoidal [41] shapes, are examples of this family of algorithms, which focus on the idea of mapping a set of 3D face models onto the images, based on a group of 2D-3D correspondences. Textured triangular meshes [28] or cubic polynomials [45] can be used in such mapping. In this model-driven family, the work of Krinidis *et* al. [26] shares some insight with the algorithm proposed in this paper, specifically by inferring the equations that govern the face deformation model, fed by the tracking module.

Discriminative models are usually holistic, and consider the whole image of the head / face for estimation, instead of local landmarks. Li *et* al. [22] estimated local image gradients, reduced dimensionality by an analysis of principal components and used a support vector regression machine to infer poses. Other similar approaches used manifold embedding algorithms (e.g., [43]) and non-linear regression methods (e.g., based on convolution networks [31]). A representative approach in this family is the work of Huang and Trivedi [14], who used a skin-tone edge-based detector to feed a tracker module based on Kalman filter and a hidden Markov model to infer poses.

Refer to the surveys published by Murphy-Chutorian and Trivedi [5], Ba and Odobez [3] and Zhang and Gao [46] for detailed information about head pose estimation and its taxonomy.

III. Proposed Method: Learning Phase

For comprehensibility, we use the following notation: matrices are represented by capitalized bold fonts and vectors appear in bold. The subscripts denote indexes. All vectors are column-wise. The ring symbol (e.g., $\mathring{\mathbf{x}}$) denotes 2D (image) positions, while 3D positions in the Euclidean space appear in regular font (e.g., $\mathbf{x}$). The hat symbol (e.g., $\hat{\mathbf{x}}$) denotes an estimate and all the hard thresholds are denoted by the κ symbol.

A. Generation of Synthetic 3D Head Shape Models

Young [44] reported 22 head dimensions from a random, composite of females and males in an adult population. The author claims these dimensions are able to describe the essential morphological properties of a human head, with 17 of these also being considered in previous surveys (e.g., [12]). Based on data from 195 females and 172 males, this study provides a set of summary statistics (minimum, maximum, mean, standard deviation, coefficient of variation, symmetry and kurtosis) for every type of measurement. In most cases, the landmarks are internal bone features, with paired surface landmarks defining lines in planes from which perpendicular distances are taken. The leftmost part of Fig. 2 illustrates some of the dimensions provided in this survey, while Table I lists the types of lengths we consider in this paper (at left) and their levels of linear correlation (rightmost matrix).

We generate the 3D head shape models randomly, starting from a single mesh that is iteratively deformed, according to

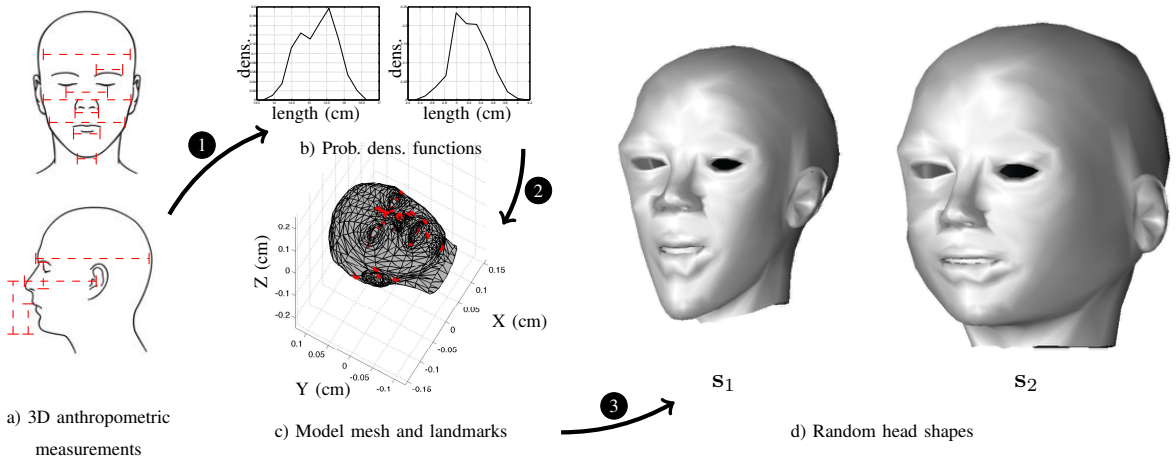
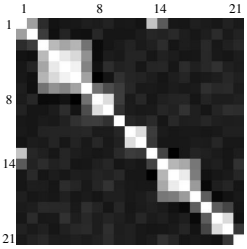


Fig. 2. Overview of the stochastic process that generates an arbitrary number of 3D head shapes (meshes). Based on anthropometric surveys (marker 1), a set of probability density functions for head lengths is defined (marker 2), and used to iteratively deform a *base* mesh, enabling to obtain head shapes of evidently different appearance (marker 3).

TABLE I
TYPES OF ANTHROPOMETRIC MEASUREMENTS CONSIDERED IN THIS PAPER AND THEIR LEVELS OF LINEAR CORRELATION.

Measurements	**Pearson Correlation**
Cranium: {1- Head circumference; 2- Head breadth; 3- Head length; 4- Biorbital breadth; 5- Biectocanthus breadth; 6- Bipupil breadth; 7- Nasal bridge breadth; 8- Bialar breadth; 9- Bicheilion breadth; 10- Bitragion breadth; 11- Bizygomatic breadth; 12- Bigonial breadth}; **Face**: {13- Sellion-menton length; 14- Sellion-supramentale length; 15- Sellion-stomion length; 16- Sellion-subnasion length}; **Nose**: {17- Midnasal bridge height; 18- Pronasale height (Maxillolaare plane); 19- Pronasale height (Sellion-promentale plane); 20- Sellion height (medial cants plane); 21- Sellion height (lateral orbital plane)}	

the target distances between the pairs of vertices. Let $\mathbf{x}_i$ be one 3D vertex and $\mathbf{n_i}$ the normal to the surface at that point. Let $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$, $\mathbf{n}_{ij} = \mathbf{n}_i - \mathbf{n}_j$ ($\mathbf{x}_\cdot, \mathbf{n}_\cdot \in \mathbb{R}^3$) and let l_{ij} be the target length (Euclidean distance) between $\mathbf{x}_i$ and $\mathbf{x}_j$. The goal is to find the magnitude of displacement α_{ij} on both vertices with respect to their normal vectors ($\mathbf{x}^{\text{new}} = \mathbf{x}^{\text{old}} + \alpha\mathbf{n}$), such that the resulting distance l_{ij} follows the probability density functions reported in the anthropometric head survey:

$$||\mathbf{n}_{ij}^T\mathbf{x}_{ij}\ \alpha_{ij}^2 + 2\mathbf{x}_{ij}^T\mathbf{n}_{ij}\ \alpha_{ij} + \mathbf{x}_{ij}^T\mathbf{x}_{ij} - l_{ij}||_2 = 0, \quad (1)$$

being $||.||_2$ the $\ell - 2$ norm. Rearranging (1) in matrix form we have:

$$||[\mathbf{n}_{ij}^T\mathbf{x}_{ij}, 2\mathbf{x}_{ij}^T\mathbf{n}_{ij}, \mathbf{x}_{ij}^T\mathbf{x}_{ij}]\ [\alpha_{ij}^2, \alpha_{ij}, 1]^T - l_{ij}||_2 = 0, \quad (2)$$

which represents one constraint of the head shape model. Let $\mathbf{c}_{ij} = [\mathbf{n}_{ij}^T\mathbf{x}_{ij}, 2\mathbf{x}_{ij}^T\mathbf{n}_{ij}, \mathbf{x}_{ij}^T\mathbf{x}_{ij}]$, $\boldsymbol{\alpha}_{ij} = [\alpha_{ij}, \sqrt{\alpha_{ij}}, 1]^T$. $\mathbf{C}$ is the block diagonal matrix that yields from the concatenation of all $\mathbf{c}$ elements, while $\boldsymbol{\alpha}$ and $\mathbf{l}$ concatenate the remaining terms:

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_{ij} & 0\,0\,0 & \dots & 0\,0\,0 \\ 0\,0\,0 & \mathbf{c}_{i'j'} & \dots & 0\,0\,0 \\ \vdots & \vdots & \vdots & \\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0 & \mathbf{c}_{i''j''} \end{bmatrix} \Bigg\} n \times 3n$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_{ij} \\ \boldsymbol{\alpha}_{i'j'} \\ \vdots \\ \boldsymbol{\alpha}_{i''j''} \end{bmatrix} \Bigg\} 3n \times 1, \mathbf{l} = \begin{bmatrix} l_{ij} \\ l_{i'j'} \\ \vdots \\ l_{i''j''} \end{bmatrix} \Bigg\} n \times 1, \quad (3)$$

being the unknowns $\boldsymbol{\alpha}$ found by:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} (\mathbf{C}\boldsymbol{\alpha} - \mathbf{l})^T\ (\mathbf{C}\boldsymbol{\alpha} - \mathbf{l}),$$
$$\text{s.t. } |||[\alpha_{ij}, \dots, \alpha_{i''j''}]|||_\infty \leq \kappa_1, \quad (4)$$

where κ_1 avoids anatomically bizarre solutions and guarantees that the solution closest to the initial configuration is preferred in the quadratic system ($\kappa_1 \approx 0.1$ in our experiments). According to this formulation, (4) is a constrained optimization problem with inequality constraints that can be solved as described in [4]. Once the $\hat{\boldsymbol{\alpha}}$ values are found, the coordinates of the corresponding vertices are updated, with similar distortions (weighted by a Gaussian kernel) applied to neighbouring vertices to enforce smoothness in the resulting mesh. The rightmost images in Fig. 2 are examples of the different meshes that can result from this stochastic process.

B. Head Shape Hypotheses

Let $\mathbf{s} = [\mathbf{x}_1^T, \dots, \mathbf{x}_{t_v}^T]^T$ be a vector representing one head shape, given as a triangulated mesh of 3D vertices. $\mathbf{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_{t_m}\}$ is the set of meshes used for learning purposes, generated as described in Sec. III-A. Evidently, there is some correlation between the $\mathbf{x}_i$ elements in each mesh, which can be attenuated by representing meshes in the principal components (PC) space:

$$\mathbf{s}^* = (\mathbf{s} - \mathbf{s}_0)\mathbf{T}_{pc}, \tag{5}$$

being $\mathbf{s}_0$ the $3t_v$-dimensional mean of the elements in $\mathbf{S}$ and $\mathbf{T}_{pc}$ the PC transformation matrix. This way, it is possible to describe each mesh in a feature space of a much lower dimension than the $3t_v$, which is important for the sake of computational effectiveness. In our case, the head models have $t_v = 957$, with 50 PC coefficients being able to represent over 99.9% of their variability.

Let $\mathbf{S}^*$ represent the shape hypotheses in the PC space. The next step is the inference of a set of prototypes that intrinsically represent *head shape similarity*, with self-organizing maps [20] (SOMs) of size $t_c \times t_c$ being considered a good choice for the following reasons: 1) SOMs obtain an ordered mapping between the 50D input space and a 2D output space, where each element represents one head prototype; 2) prototypes in the output space are topologically ordered, i.e., neighbor prototypes feature similar head shapes; 3) SOM prototypes reflect the variations in density in the input space, i.e., densely populated regions in the input space (where the most frequent head shapes fall) are represented by the largest number of prototypes; and 4) SOMs are known to be particularly suitable to model non-linear input spaces, such as our input feature space. In practical terms, the SOM output space is a similarity graph, which is important in order to label degraded data: even if a query is not mapped directly to the same cell as the enrolment sample with a corresponding identity, it should be mapped to a neighboring cell. Fig. 3 illustrates the head prototypes (cells) that are used as soft labels.

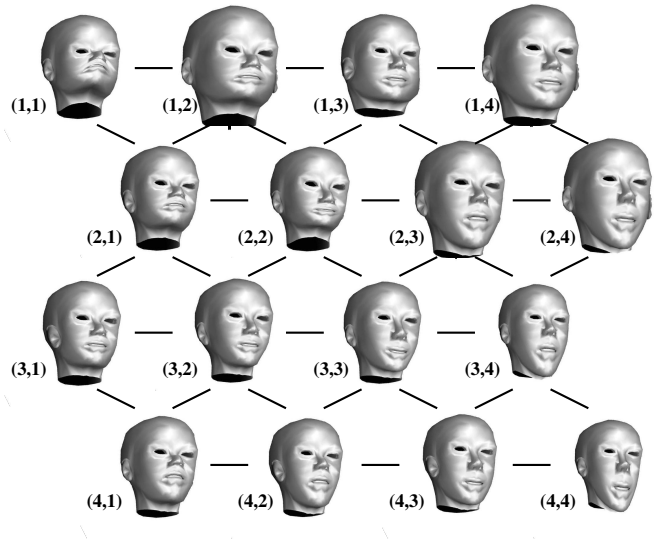
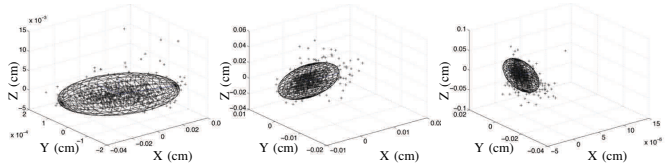


Fig. 3. Representation of the 3D head centroids resulting of a 4×4 SOM. Note the similarity in size / shape between adjacent elements, rooted in the preservation of the topological properties of the input space that this kind of maps offers.

C. 3D Head Shape Covariance

Let $\mathbf{s}_{c_i}$ be the head shape centroid corresponding to the i^{th} cell in the SOM, and let $\{\mathbf{s}_{c_{i1}} \ldots, \mathbf{s}_{c_{iw}}\}$ be the shape samples associated with $\mathbf{s}_{c_i}$. For all the elements in $\mathbf{s}_{\cdot}$ that correspond to head landmarks, the displacement between the 3D positions in the samples and in the centroid were measured $(\mathbf{x}_{c_{ij}} - \mathbf{x}_{c_i})$, obtaining a set of 3D vectors from where the mean and covariance matrix were taken. This captures the spread of the 3D data and is used in the algorithm convergence test to discriminate between genuine / spurious query landmarks. To illustrate this point, Fig. 4 plots the 99% confidence ellipsoids for the *right ear lobe*, *center of right cornea* and *nose apex* landmarks.



a) Right ear lobe b) Center of right cornea c) Nose apex

Fig. 4. Examples of the 99% confidence ellipsoids that represent the deviations of the positions of landmarks in the head shape samples with respect to their centroid. These values are used in the convergence test of the algorithm to discriminate between genuine / spurious head landmarks.

D. Pose Hypotheses

Let $\mathbf{p} = \{\mathbf{R}, \mathbf{t}\}$ be a camera pose configuration, with $\mathbf{R}$ being the rotation matrix and $\mathbf{t}$ the translation vector, i.e., $\mathbf{p}$ is a 6D vector accounting for three components of rotation (yaw, pitch and roll) and three of translation (t_x, t_y and t_z). Let $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_{t_p}\}$ be a set of pose hypotheses, created randomly using uniformly distributed random numbers for all six degrees of freedom. Given the relatively large number of elements generated ($\approx$ 100,000), a set of pose prototypes is also obtained. In this case, as there are no requirements about the concept of *similar* poses, such prototypes can be found simply by the k-means algorithm, yielding $t_{\tilde{p}}$ pose vectors ($t_{\tilde{p}} \ll t_p$).

E. Joint Head Shapes / Pose Hypotheses Indexing

Given a set of $t_{\tilde{p}}$ pose and t_c^2 head shape hypotheses, during classification it is required to find the *best* joint pose / shape configuration, which is the most likely match to the query. Theoretically, there are a total of $t_{\tilde{p}} t_c^2$ possibilities, but exploring all by brute-force is prohibitive in terms of time complexity. Moreover, not all the query landmarks will be genuine, and both false negatives and false positives are expected. Given such constraints, a forest of binary trees was created, one per type of landmark, where the hypotheses are grouped (k-means) in leaves according to their neighborhood of one landmark projection, given by the *world-to-image* function:

$$f_{w \to i}(\mathbf{x}, \mathbf{p}) = \frac{1}{\upsilon}\mathbf{A}[\mathbf{R}|\mathbf{t}] \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}, \tag{6}$$

being $\mathbf{x}$ the vertices of $\mathbf{s}$, υ the scalar projective parameter, $\mathbf{A}$ the internal camera matrix, and $\mathbf{p}$=$\{\mathbf{R}$ (rotation), $\mathbf{t}$ (translation)$\}$ the pose parameters. This way, each tree keeps, within

its leaves, the indices of the hypotheses that have similar 2D projections of a landmark. Later, in classification, the position of every query landmark is used in the corresponding tree to obtain the indices of the complying hypotheses. By repeating the process for all landmarks and accumulating the complying indices, the hypotheses are ranked in descending order according to the frequency with which they appear in leaves, so that the most likely (those with the highest number of landmarks close to the query) will be evaluated first.

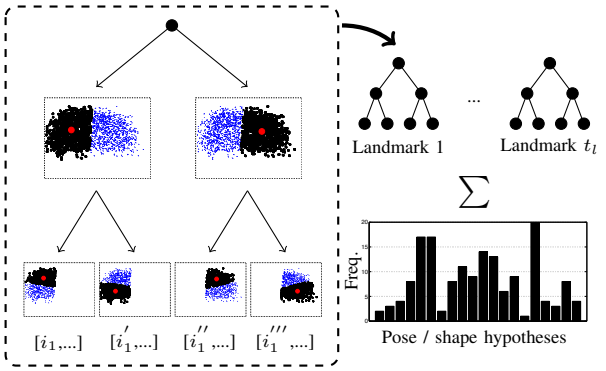


Fig. 5. Data structure that indexes the joint pose / shape hypotheses, grouped according to the similarity of their landmark projections. In retrieval, the indices of the hypotheses complying the query landmarks are accumulated, such that the most voted hypotheses will be evaluated first.

The retrieval process is illustrated in Fig. 5, and has a time complexity $\mathcal{O}(t_l \log(t_{\tilde{p}} t_c^2))$, t_l being the number of query landmarks. This roughly logarithmic time complexity is important for generating large sets of hypotheses without substantially compromising the time cost of retrieval.

IV. Classification Phase

Let $\mathring{\mathbf{q}} = \{\mathring{\mathbf{q}}_1, \ldots, \mathring{\mathbf{q}}_{t_q}\}$ be a set of 2D head landmarks in a query image. We assume that the *type* of each landmark $\tau(\mathring{\mathbf{q}}_i)$ is known, i.e., the anatomic region corresponding to each $\mathring{\mathbf{q}}_i$ is given as input. This is a readily satisfied assumption, using the state-of-the-art techniques for head / face landmark detection (e.g., [18], [8], or [33]).

Using the trees described in Sec. III-E, the most likely joint pose / head shape hypothesis for the query is obtained and its pose configuration subsequently optimized. Assuming that the pose hypothesis $\mathbf{p}$ is *relatively* close to the query pose, the idea is to perform only small adjustments in its parameterization to better fit the query:

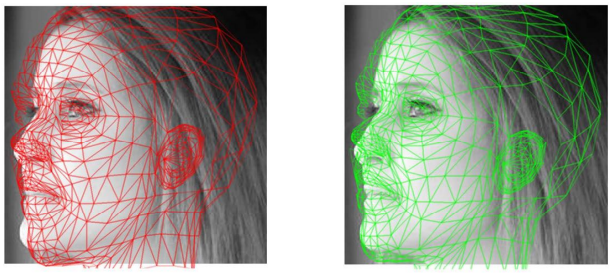
$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} d\Big(f_{w\to i}(\mathbf{s}, \mathbf{p}), \mathring{\mathbf{q}}\Big), \tag{7}$$

where $f_{w\to i}(\mathbf{s}, \mathbf{p}) = f_{w\to i}(\mathbf{x}, \mathbf{p}), \forall \mathbf{x} \in \mathbf{s} = \mathring{\mathbf{s}}$ and $d(.,.)$ is the function that measures the similarity between two sets of landmarks:

$$d(\mathring{\mathbf{s}}, \mathring{\mathbf{q}}) = \frac{1}{\nu(\mathring{\mathbf{q}})} \sum_{i=1}^{\nu(\mathring{\mathbf{q}})} \min_{\mathring{q}_j | \tau(\mathring{q}_j)=\tau(\mathring{\mathbf{x}}_i)} d(\mathring{\mathbf{x}}_i, \mathring{\mathbf{q}}_j), \tag{8}$$

where $d(\mathring{\mathbf{x}}, \mathring{\mathbf{q}}) = ||\mathring{\mathbf{x}} - \mathring{\mathbf{q}}||_2$ and $\nu(\mathring{\mathbf{q}})$ is the function that counts the number of distinct types of landmarks in $\mathring{\mathbf{q}}$. Essentially, (8) sums the distances between projections of 3D head vertices and their closest query landmarks of the corresponding type.

The optimization process is regarded as convex and unconstrained, with all the advantages inherent to it in terms of computational cost. We use a derivative-free algorithm proposed by Lagarias *et* al. [21], due to its proven effectiveness in relatively low dimensionality problems (six in our case). Having an initial pose hypothesis $\mathbf{p}$, the algorithm generates a sample of seven points around $\mathbf{p}$ and iteratively discards the point with the maximum value of the cost function (8), replacing it with a new point generated either by reflection, expansion, contraction or shrinkage of sample points. As Fig. 6 illustrates, this process enables us to better fit the pose hypothesis to the query data by only slightly adjusting the initial configuration.



a) Initial pose / shape hypothesis $\mathbf{p}$ b) Shape with optimized pose $\hat{\mathbf{p}}$

Fig. 6. Pose refinement, according to a convex optimisation paradigm. Assuming that the initial hypothesis $\mathbf{p}$ is a good approximation of the solution, the probability of falling in local minima is relatively short. $\hat{\mathbf{p}}$ is the optimized configuration.

Having an optimized estimate of pose $\hat{\mathbf{p}}$, the final step is the evaluation of the reasonability of the $\{\hat{\mathbf{p}}, \mathbf{s}\}$ solution, either stopping the algorithm or continuing to the next hypothesis. This evaluation is carried out in the 3D space by inferring the most likely 3D positions for the query landmarks. Let $\mathring{\mathbf{q}} = (x, y)$ be one image landmark corresponding to one vertex in $\mathbf{s}$. There is a ray in the Euclidean 3D space from where elements are projected into $\mathring{\mathbf{q}}$, which is given by the *image-to-world* function:

$$f_{i\to w}(\mathring{\mathbf{q}}, \hat{\mathbf{p}}) = \mathbf{R}^T \mathbf{A}^{-1} \upsilon \begin{bmatrix} \mathring{\mathbf{q}} \\ 1 \end{bmatrix} - \mathbf{R}^T \mathbf{t}, \tag{9}$$

with $\mathbf{A}$ being the internal camera parameters, $\mathbf{R}$ and $\mathbf{t}$ its extrinsic parameters (obtained from $\hat{\mathbf{p}}$) and υ being the scalar projective parameter. The shortest distance between the ray and the corresponding vertex in $\mathbf{s}$ is the most *optimistic* location of $\mathring{\mathbf{q}}$ in the 3D space:

$$\hat{\mathbf{q}} = \mathbf{x}_r + \mathbf{v}_r^T \odot \frac{(\mathbf{x} - \mathbf{x}_r)^T \mathbf{v}_r}{||\mathbf{v}_r||^2}, \tag{10}$$

being $\odot$ the point-by-point multiplication operator, $\mathbf{x}_r, \mathbf{v}_r$ the 3D point and vector defining the ray (given by (9)). Fig. 7 illustrates the rationale behind this step, where the 3D positions $\hat{\mathbf{q}}_{\cdot}$ from where the query landmarks $\mathring{\mathbf{q}}_{\cdot}$ might have been projected are estimated based on $\{\hat{\mathbf{p}}, \mathbf{s}\}$.

According to (10), only the query landmarks $\hat{\mathbf{q}}^*$ that are the most likely to be genuine are selected, providing the minimum

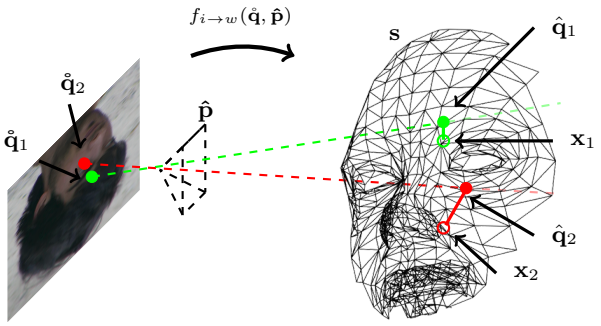


Fig. 7. Finding the 3D positions in the Euclidean space from where the query landmarks might have been projected, according to a pose $\hat{\mathbf{p}}$ and shape $\mathbf{s}$ estimates. The $||\hat{\mathbf{q}}_\cdot - \mathbf{x}_\cdot||_2$ values are used to discriminate between the spurious (in red) and genuine (in green) query landmarks.

$||\hat{\mathbf{q}}_\cdot - \mathbf{x}_\cdot||_2$ values (per type of landmark). Henceforth, all the remaining landmarks are deemed to be spurious and are discarded. Finally, given the set of remaining landmarks and their most plausible 3D positions, ϕ evaluates the reasonability of such positions by checking if misalignments are inside the prediction interval ellipsoid, obtained as described in Sec. III-C:

$$\phi(\hat{\mathbf{q}}_i^*|\mathbf{x}_i, \mathbf{x}_{c_i}, \mathbf{\Sigma}_i) = (\hat{\mathbf{q}}_i^* - \mathbf{x}_i - \mathbf{x}_{c_i})^T \mathbf{\Sigma}_i^{-1} (\hat{\mathbf{q}}_i^* - \mathbf{x}_i - \mathbf{x}_{c_i}) - \chi_3^2(0.99), \tag{11}$$

with $\mathbf{x}_{c_i}$ as the position of the shape centroid, $\mathbf{\Sigma}_i$ as the covariance matrix and $\chi_3^2(0.99)$ as the quantile function for probability 99% of the chi-squared distribution with three degrees of freedom. In practical terms, this function checks if it is likely to observe a $\hat{\mathbf{q}}_\cdot^* - \mathbf{x}_\cdot$ misalignment between a sample landmark and its centroid, returning a positive value if the misalignment falls inside the covariance error ellipsoid (Fig. 4) and a negative value otherwise. Finally, a solution is *acceptable* if a sufficient number of landmarks is deemed genuine, i.e., $H_{all}() \geq \kappa_2$:

$$H_{all}(\hat{\mathbf{q}}^*|\mathbf{x}, \mathbf{x}_{c_i}, \mathbf{\Sigma}) = \frac{1}{\nu(\hat{\mathbf{q}}^*)} \sum_{i=1}^{\nu(\hat{\mathbf{q}}^*)} H\big(\phi(\hat{\mathbf{q}}_i^*|\mathbf{x}_i, \mathbf{x}_{c_i}, \mathbf{\Sigma}_i)\big), \tag{12}$$

where κ_2 is the convergence threshold, $\nu(\hat{\mathbf{q}}^*)$ is the number of query landmarks, and H is the Heaviside function:

$$H(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{if } z < 0. \end{cases} \tag{13}$$

V. Results and Discussion

Three well known data sets were selected for our experimental evaluation. The Annotated Facial Landmarks in the Wild [19] (AFLW) set was used to evaluate the results of the pose estimation phase. It has 25,993 color images, each one annotated with a 21-point markup on visibility. In this set, we considered exclusively samples with pose angles in the intervals yaw $\pm\pi/4$, pitch $\pm\pi/2$, and roll $\pm\pi/5$, according to the plausibility of observing such poses in visual surveillance scenarios. The soft biometric labels were evaluated using the Labeled Faces in the Wild [15] (LFW) and in the SCface [10] sets, selected due to the wildness of their data. Out of the 9,164 images in the LFW set, 670 were disregarded due to extremely poor performance of the head landmark detector, resulting in 8,494 samples from 1,574 subjects. For the SCface set, we exclusively used the third sample from cameras 1-5 (650 images from 130 subjects), which have the maximal resolution acquired at visible wavelengths. Fig. 8 shows some images from the data sets considered. In all the experiments below, the thresholds were set to $\kappa_1 = 0.01$ and $\kappa_2 = 0.9$.

Fig. 8. Examples of the data sets used in the empirical validation of the proposed method. The upper row regards the AFLW data set, whereas the bottom rows are from the LFW and SCface sets.

A. Pose Estimation

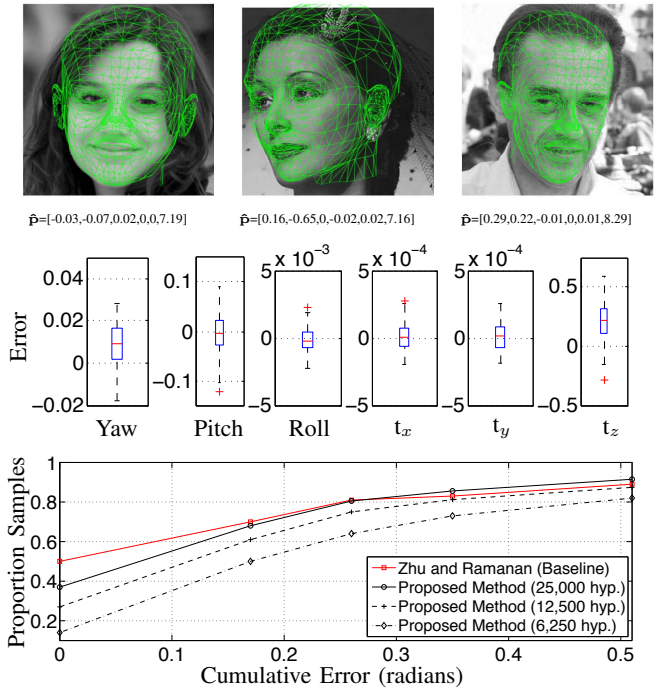


Fig. 9. Upper row: examples of pose estimates in images from the AFLW data set. Second row: boxplot of the pose estimation errors for the six degrees-of-freedom: {yaw, pitch, roll} rotation angles (in radians), plus the $\{t_x, t_y, t_z\}$ translation values. Bottom row: performance comparison with respect to a state-of-the-art pose estimator [48] in a subset of the AFLW set.

Let $\mathbf{p} \in \mathbb{R}^6$ be the ground-truth pose of a sample and $\hat{\mathbf{p}}$ be the pose configuration found by our algorithm. In Fig. 9,

we give the box plots of the $\mathbf{p} - \hat{\mathbf{p}}$ values for each of the 6 pose degrees of freedom, showing the median of the errors (horizontal solid line) and their first and third quartile values (top and bottom of the box marks). The upper and lower whiskers are denoted by the horizontal lines outside each box, and the outliers are denoted by crosses. The upper row exemplifies three queries and the corresponding poses found by the algorithm. In these experiments we used 25,000 joint poses / head shape hypotheses, i.e., $t_{\tilde{p}} = 1,000, t_c^2 = 25$, indexed in binary trees of height 10 ($\approx$ 50 hypotheses per leaf).

Overall, upon the availability of a sufficient number of pose hypotheses, the algorithm obtained a visually pleasant approximation of the query poses for the large majority of the cases. Objectively, we compared the performance of our pose estimator, with 6,250, 12,500 and 25,000 joint head shapes / pose hypotheses ($t_{\tilde{p}} = \{250, 500, 1,000\}$, $t_c^2 = 25$), to a state-of-the-art method due to Zhu and Ramanan [48], using the data set these authors supply[1]. The cumulative error curves are given in the bottom plot of Fig. 9, with the best configuration in our solution attaining performance close to the state-of-the-art, but using a much lower (and unfiltered) number of facial landmarks than the baseline. Overall, the gap in performance between both methods was the largest for low error values (where a larger number of landmarks would be particularly useful), and the results tended to converge for large cumulative errors which correspond to rough pose estimates. For large cumulative errors - over $\frac{\pi}{4}$ - our method (with $t_{\tilde{p}} = 1,000$) attains better pose estimates than the baseline. We note that errors increase substantially when a reduced number of pose hypotheses are generated, particularly for $t_{\tilde{p}}$ values below 250. However, it should also be noted that generating large sets of hypotheses in the learning phase is not a concern, as the indexing strategy used accounts for several thousands of hypotheses without significantly increasing the temporal complexity of retrieval.

B. Soft Labels' Stability

The stability of the proposed soft labels varies per subject and depends of the number of SOM centroids. We define the stability of the ith subject as:

$$S_{t_c}(i) = 1 - \frac{1}{\sqrt{2} t_c t_i} \sum_{a=1}^{t_i} ||\mathbf{b}_{ia} - \bar{\mathbf{b}}_i||_2, \qquad (14)$$

with $\mathbf{b}_{ia} \in \mathbb{N}^2$ being the ath sample label for the ith subject, $\bar{\mathbf{b}}_i$ being the subject centroid label ($\bar{\mathbf{b}}_i = \frac{1}{t_i} \sum_{a=1}^{t_i} \mathbf{b}_{ia}$), t_i being the number of samples of the subject and t_c denoting the number of columns / rows in the SOM (only square SOMs were considered).

For a set of subjects, a summary of their stability is given by $S_{t_c} = \frac{\sum_{i=1}^{t_s} S_{t_c}(i)\ t_i}{\sum_{i=1}^{t_s} t_i}$, t_s being the number of subjects. Fig. 10 depicts the stability of labels in the LFW data set, with respect to the number of centroids. The left plot gives three probability density functions for the $S_{t_c}(i)$ values using three typical SOM sizes. The right plot gives the group stability S_{t_c}, again as function of the SOM size.

[1] http://www.ics.uci.edu/~xzhu/face/

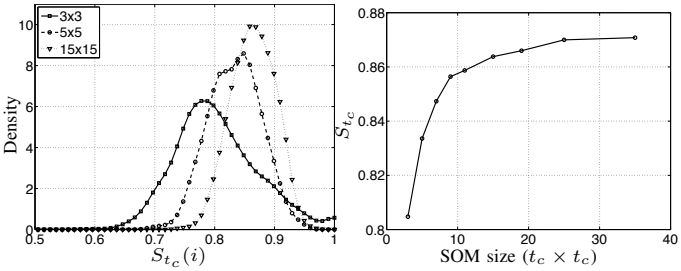


Fig. 10. Left: Probability density functions of the stability of labels per subject ($S_{t_c}(i)$). Right: Variations in the overall stability S_{t_c} with respect to the number of shape centroids considered.

The $S_{t_c}(i)$ values varied from around 0.63 (worst case for small maps) to 1 in the LFW set, with the optimal value observed for subjects with head shapes associated with cells in a SOM corner. Also, by using small SOMs (e.g., 3 $\times$ 3) the probability of obtaining near optimal stability values (all samples of a subject associated to the same cell) is increased, but so is also obtaining many more low stability values. Note that in small maps even small misalignments correspond to large normalized distances.

Overall, the summary stability S_{t_c} varied in direct correspondence with the number of cells in the SOM, converging for values around 0.87 in maps with more than 20 $\times$ 20 cells. Note that (14) provides relative distances with respect to the size of the SOM, i.e., values equal to 1 occur when two labels are separated by $\sqrt{2}\ t_c$ (a SOM diagonal). This explains why the stability values increase for larger SOMs, even though small maps should intuitively provide the maximum stability.

C. Soft Labels' Discriminability

The discriminability of labels was evaluated based on the flatness of the histogram that counts the number of subject centroids per cell, considering that discriminating labels should spread subjects evenly across the SOM cells. This is measured by an entropy function:

$$D_{t_c} = -\sum_{i=1}^{t_c^2} p_i \log_{t_c^2} p_i, \qquad (15)$$

where p_i is the empirical probability that a subject centroid is associated with the ith cell of the SOM. In this case, the subject centroid labels were rounded to their closest cell. Being $D_{t_c} \in [0, 1]$, values close to 1 denote flat histograms, where subjects are spread evenly across the SOM cells. Values close to 0 are the non-interesting case, where most subjects are associated with a reduced number of cells.

Fig. 11 expresses the D_{t_c} values with respect to the SOM dimensions, having attained a maximum for the smallest maps (3 $\times$ 3), with an approximately equal number of subject centroids per cell. As the number of cells increased, some of the cells started to have too few centroids, while others attracted the elements in that region, yielding a more uneven distribution of the number of elements per cell.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2522441, IEEE Transactions on Pattern Analysis and Machine Intelligence

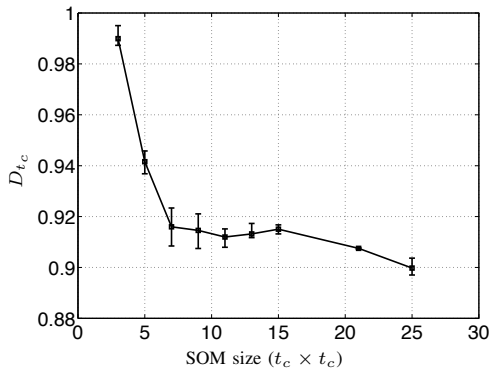


Fig. 11. Relationship between the labels' discriminability and the dimension of the SOMs used.

Fig. 12 gives examples of the associations between the queries and the 3D head centroids for the LFW data set, using a 10 $\times$ 10 SOM. In each row, the leftmost image is the 3D head shape centroid (label) and the remaining images illustrate samples associated with that cell. Note the evident similarity between the major head features of the subjects and the centroids: at the upper-left extreme in the SOM, the (1,1) cell represents the largest heads with a round shape. At the other extreme, the (10,10) cell represents the most longitudinal heads with salient chins and extent maxillae. In this kind of mapping, cells in the corners provide the most easily distinguishable features (under visual inspection), while central cells are not so obviously distinguishable with respect to neighbors (note the high similarity between elements in cells (3,2) and (3,3)). Moreover, as the central region represents the most densely populated region of the feature space, a larger number of prototypes is used here, which accounts for the higher similarity between neighbouring prototypes.

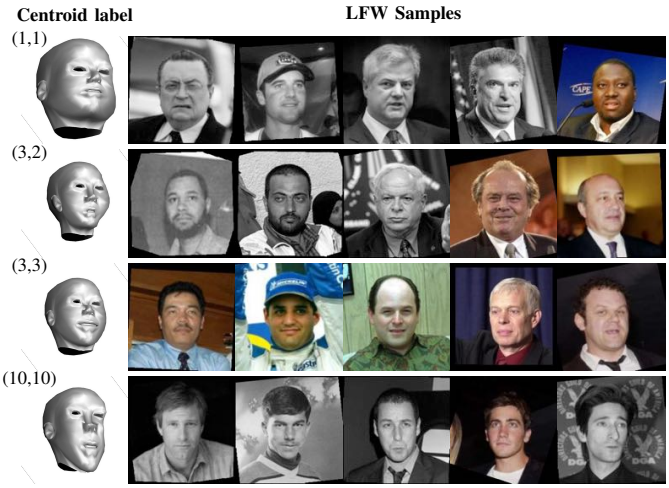


Fig. 12. Examples of associations between samples of the LFW data set and the head shape centroids of a SOM with 10 $\times$ 10 cells.

D. Robustness to Clutter

Poor-quality data queries are expected to be cluttered, i.e., with misplaced landmarks not corresponding to the anatomical region they are supposed to represent. This section addresses the effects of such cluttered input in the algorithm performance, which are two-fold: 1) increase the number of head shapes / pose hypotheses explicitly evaluated before convergence; and 2) decrease the convergence rate of the algorithm, which occurs when a solution is not found after evaluating the maximum number of hypotheses (100 in our experiments). Let p_s be the proportion of spurious landmarks with respect to the accurate detections (e.g., $p_s = 0$ represents a non-cluttered input and $p_s = 1$ denotes a balanced number of spurious / genuine landmarks). Using images of the AFLW set (with landmarks confirmed by human observers), cluttered inputs were simulated, by adding landmarks away from their true position (random x, y coordinates uniformly distributed over the entire image space, $\mathcal{U}(0,1)$, with coordinates normalized in the [0,1] interval) or by changing the position of a landmark (again, by generating uniformly distributed displacements over the image space, $\mathcal{U}$(0,1)).

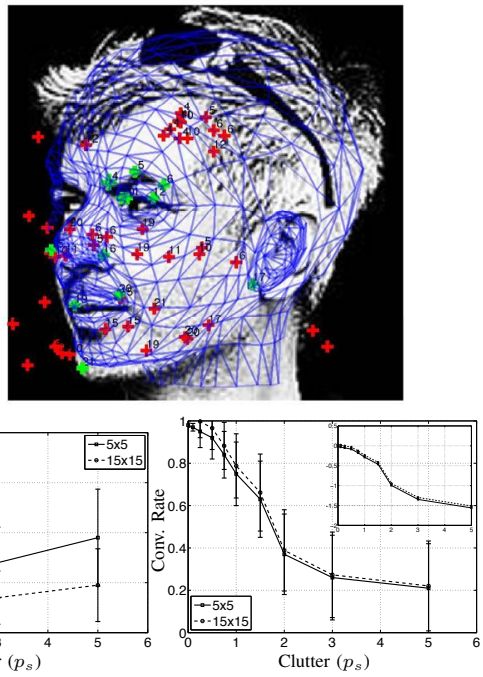


Fig. 13. Top: illustration of a query sample with spurious head landmarks (p_s=4), where the proposed method was still able to correctly estimate the pose and the soft biometric label. Bottom-left plot: effect of the proportion of spurious landmarks in the number of joint pose / head shape hypotheses explicitly evaluated before convergence (given in linear and log scales). Bottom-right plot: decay in the convergence rate with respect to the proportion of spurious correspondences (linear and log scales).

As illustrated in Fig. 13, the algorithm convergence rate decayed with respect to p_s, but only slightly for values below one, which is readily achieved by state-of-the-art head landmark detectors. For larger p_s values, the convergence rate of the algorithm decays evidently and, for $p_s > 5$, the algorithm loses its effectiveness (bottom right plot). In terms of the number of

hypotheses explicitly evaluated, an approximately direct linear relationship with respect to p_s was observed (bottom left plot). The top image in Fig. 13 illustrates a query with $p_s = 4$ and the output of the algorithm, where the landmarks deemed genuine (with $\phi() \geq 0$) appear in green and the spurious landmarks are denoted by the color red.

E. Soft Labels Standalone Performance: The Watch-list Problem

An important surveillance task is the watch-list problem: authorities have an explicit list of criminals (the *watch-list*) they want to locate or track among a population. Given a query, the goal is to detect occurrences of watch-list elements without revealing the identities of any other subjects to the central authorities, which is considered a privacy-preserving policy.

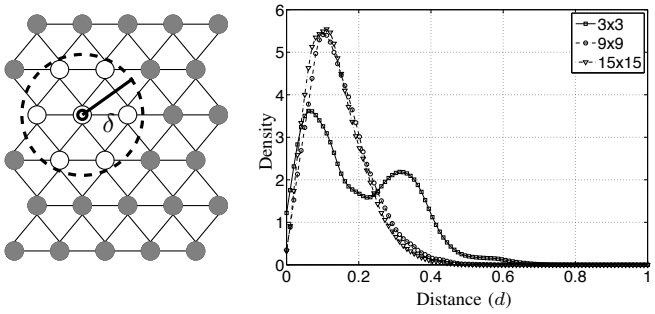


Fig. 14. At left: insight of the *negative identification* concept, used in the watch-list problem. All labels farther than δ of a query correspond to identities that can be rejected. At right: probability density functions of observing distances d between intra-subject labels (values regard the LFW data set).

The metric space formulation of labels is particularly suitable for handling this type of problem. By assigning a cell to each element in the watch-list, the topological properties of the input space ensure that any query assigned with cells located *sufficiently* far from the watch-list cell does not correspond to the criminal's identity. This is illustrated in the left diagram in Fig. 14. Depending on the radius δ used (which dictates the relationship between the hit / penetration rates), most of the identities in the watch-list can be confidently rejected. The plot given at the right side of Fig. 14 shows the probability density functions of observing distances d between intra-subject samples, which is the key for this watch-list formulation. Values are given for SOMs of three different sizes and enable us to conclude that there is a minimal probability of observing large distances (> 0.5) between intra-subject labels.

The suitability of the soft labels for watch-list identification is confirmed in the results given in Fig. 15, which expresses the hit / penetration values for the LFW set, using SOMs of dimensions 3×3 (continuous line with square marks), 9×9 (dashed line with circular marks), and 15×15 (dot-dashed line with triangular marks). The performance lines of the largest SOMs almost overlap and enable to reject over 50% of the identities for a query, keeping hit rates close to 99%.

Fig. 16 gives the hit / penetration values obtained for the SCface set, which are worse than the LFW values. This was justified by the small image resolution in the set, making the

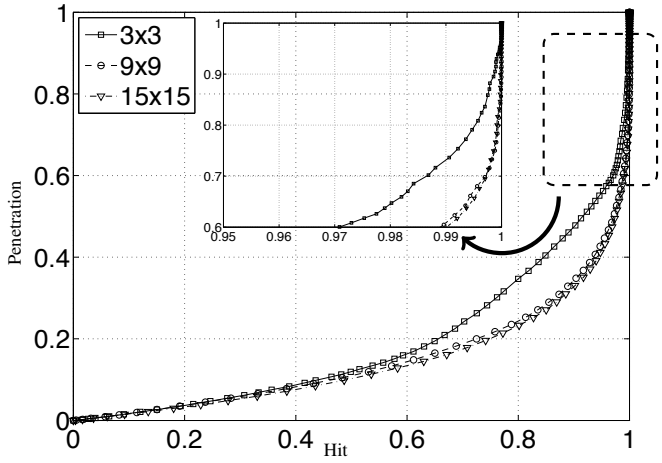


Fig. 15. Hit / penetration plots for the LFW data set, using SOMs of dimensions 3×3 (continuous line), 9×9 (dashed line) and 15×15 (dot-dashed line).

detection of head landmarks an extremely difficult task. Also, poses variations in this set are constrained to pitch angles (yaw and roll angles close to 0, pitch values in $[\pi/40, \pi/10]$), which led us to use only 100 pose prototypes. However, the key factor behind the relatively poor performance was that, in data of such reduced resolution, even small inaccuracies in landmark detection lead to large deviations in the 3D model positions inferred, which considerably reduced the stability of labels.

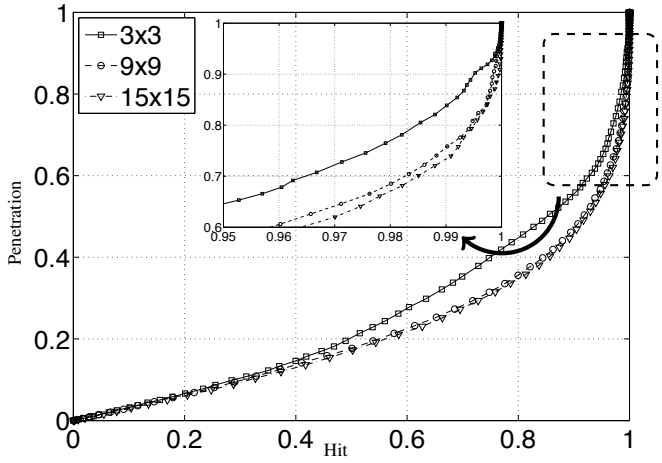


Fig. 16. Hit / penetration plots for the SCface data set, using SOMs of dimensions 3×3 (continuous line), 9×9 (dashed line) and 15×15 (dot-dashed line).

F. Fusion of Soft / Strong Traits: Recognition Performance

This section addresses the effectiveness of the soft labels to provide auxiliary information to a strong biometric expert. As in the previous sections, the LFW was used as main data set, having chosen the evaluation mode (*unsupervised*) that provides the lowest recognition performance among all protocols[2]. As

[2]http://vis-www.cs.umass.edu/lfw/results.html

a baseline, we considered the face recognition method due to Arashloo and Kittler [2] based on two reasons: 1) this method is among the best performers in the unsupervised (training free) LFW evaluation mode; and 2) it integrates well known techniques in a typical biometric recognition processing chain that could be easily applied to other traits (i.e., the ocular or the ear regions). It uses a multi-layered graphical model that measures the geometric distortion between image pairs, fed by the Daisy [39] feature descriptor. In classification, multi-resolution LBPs, image registration techniques and the cosine similarity yield the pairwise similarity score.

Note that the purpose of these experiments is not to obtain a system that outperforms the face recognition state-of-the-art, but to show that the proposed type of weak trait can be fused with strong systems and still improve the recognition performance with respect to the baseline. From this perspective, the *relative* performance between the ensemble and the baseline is most important than the absolute effectiveness rates. Also, note that other improvements in performance with respect to the baseline could be obtained by properly using the landmarks information provided by the soft expert inside the face recognition engine. However, that will be an attempt to improve a *specific* face recognizer, which is out of the scope of this paper.

The face and soft biometric experts were fused at the score level, learning a linear discriminant that projects both scores into the subspace that maximizes the Fisher discriminant ratio (found in a disjoint set composed by 10% of the available pairwise comparisons). Let ϵ_f be the pairwise similarity score returned by the face recognition expert and ϵ_s be the score returned by the soft expert:

$$\epsilon_s = \frac{1 + \text{erf}\big(\kappa(\frac{||\mathbf{b}_1 - \mathbf{b}_2||_2}{\sqrt{2}t_c} - 0.5)\big)}{2}, \qquad (16)$$

where erf() is a transfer function (error function) with sigmoid shape, $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{N}^2$ are the labels (of a $t_c \times t_c$ SOM) associated to the image pairs and κ is the parameter that controls the shape of the transfer function ($\epsilon_s \in [0, 1]$). Results are summarized in the Receiver Operating Characteristic curves of Fig. 17: the black line gives the baseline performance of the face expert, and the colored lines are the results attained by the ensemble, for three different shapes of transfer functions ($\kappa \in \{1, 2, 4\}$), with larger values corresponding to those farther from linear shapes. When compared to the baseline, the improvements in performance were maximized when the transfer function had the most pronounced sigmoid shape ($\kappa = 4$), i.e., when small misalignments between $\mathbf{b}_1$ and $\mathbf{b}_2$ were not excessively penalized. On the other hand, for roughly linear transfer functions ($\kappa \approx 1$), the performance of the ensemble was even slightly worse than the baseline.

Analyzing in detail the $\kappa = 4$ ensemble, we concluded that improvements in performance were due to reducing the variability of intra-subject scores, typically by improving the pairwise scores when both samples had largely different poses, with the face recognition expert showing a particular sensitivity to such covariates (cases where the graphical model was not able to infer the appropriate deformation parameters). Conversely, we observed that the impostors' score distributions in the baseline and in the ensemble were almost equal.

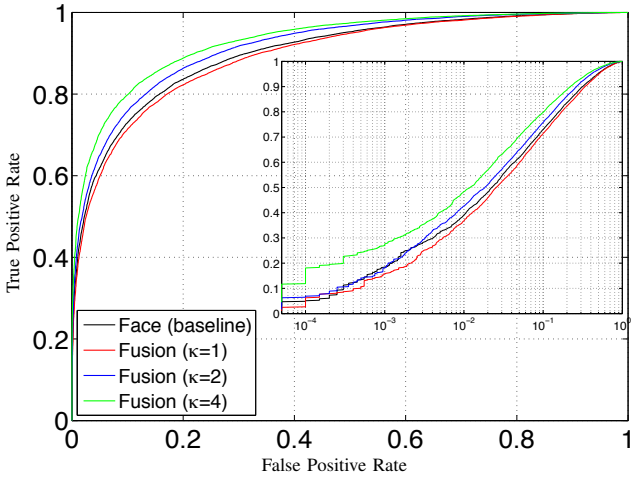


Fig. 17. Comparison between the recognition performance attained by a face recognition system in standalone mode and when using also the soft biometric labels as auxiliary information. Results are given for the LFW data set and regard the *unsupervised* evaluation mode.

G. Effect of Facial Expressions

Considering that facial expressions may significantly distort the head morphology (Fig. 18), this section addresses the effect of facial expressions on the soft labels from three perspectives. Initially, it reports the deviations in the SOM cells associated with intra-subject samples with neutral / non-neutral expressions. Next, it compares the labels' stability / discriminability in three different scenarios: 1) with 3D head shapes and queries having *neutral* expression; 2) with *neutral* 3D shapes against queries of *unconstrained* (*neutral* and *non-neutral*) expressions; and 3) with *unconstrained* 3D shapes and queries. Finally, it evaluates the suitability of the soft labels recognizing facial expressions.

All the experiments were conducted using the previously mentioned LFW set, with images divided into disjoint groups, according to the facial expressions considered. Additionally, the Extended Kohn-Canade (CK+) [24] set was selected, being one of the most popular sets in this research topic. In terms of the facial expressions considered, we constrained the analysis to the *neutral* and *happy* expressions, due to two reasons: 1) the recognition of the remaining types of facial expressions (e.g., fear, disgust or sadness) implies the detection of action units that depend of an excessively large number of facial landmarks that cannot be detected in poor quality data; and 2) the LFW set has a small number of samples with other facial expressions (apart from *neutral* and *happy*), as they are unlikely in visual surveillance scenarios. In these experiments SOMs had 15 $\times$ 15 cells, maintaining all the κ_i values used previously.

Initially, only 3D head shapes of *neutral* expression were generated, with queries grouped per individual and per facial expression. For each subject, the centroid labels for *neutral* $\bar{\mathbf{b}}_i^{(n)}$ and *happy* $\bar{\mathbf{b}}_i^{(h)}$ expressions were found. The left plot in

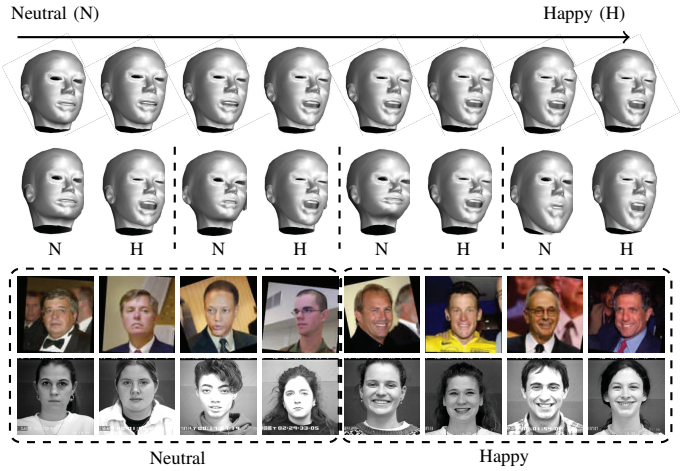


Fig. 18. Upper row: variations in the head shape appearance with respect to the levels of evidence of a facial expression (*happy*). Second row: pairwise samples of 3D head shapes with *neutral* / *happy* facial expression. Bottom row: division of the samples from the LFW and Cohn-Kanade data sets into two disjoint groups, according to their facial expression.

Fig. 19 gives the velocity plot corresponding to the $\bar{\mathbf{b}}_i^{(h)} - \bar{\mathbf{b}}_i^{(n)}$ misalignments, showing the average magnitude / direction of vectors representing the typical movements in SOM labels when expressions change from *neutral* to *happy*. It is evident that movements vary across the maps, with central regions being more stable than regions near the corners. Overall, movements converge in the bottom-right corner that represents the most elongated faces (with the largest deformations in the head shape due to the *happy* expression). The rightmost part of Fig. 19 gives two examples of *neutral* / *happy* head shapes falling in the SOM regions where the largest deviations were observed.

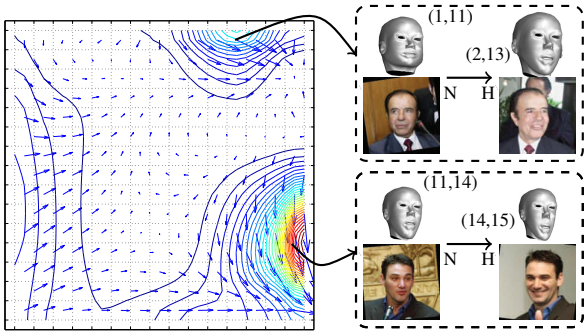


Fig. 19. At left: velocity plot representing the predominant intra-subject displacements in labels with respect to changes in facial expression from *neutral* to *happy* ($\bar{\mathbf{b}}_i^{(h)} - \bar{\mathbf{b}}_i^{(n)}$), using 3D head prototypes of exclusively *neutral* expression. At right: samples associated to the SOM regions with the largest movement slopes, i.e., where facial expressions imply the largest misalignments between the positions of soft labels in the SOM.

In addition, to perceive the decrease in soft labels effectiveness due to facial expressions, Fig. 20 compares the labels' stability / discriminability for three distinct configurations: 1) using 3D head shapes and queries exclusively of *neutral* expression; 2) using *neutral* head shapes and *unconstrained* queries (i.e., samples with *neutral* / *non-neutral* expressions); and 3) using *unconstrained* head shapes and queries. Results are given in terms of the hit / penetration plots and show that facial expressions consistently decrease the effectiveness of soft labels. However, such degradation is counter-balanced if 3D shape hypotheses with *non-neutral* expressions are also generated, yielding results that are not too far off the baseline *neutral* against *neutral* configuration (at the expense of an increase in the computational burden of the labelling task by doubling the number of head shape hypotheses).

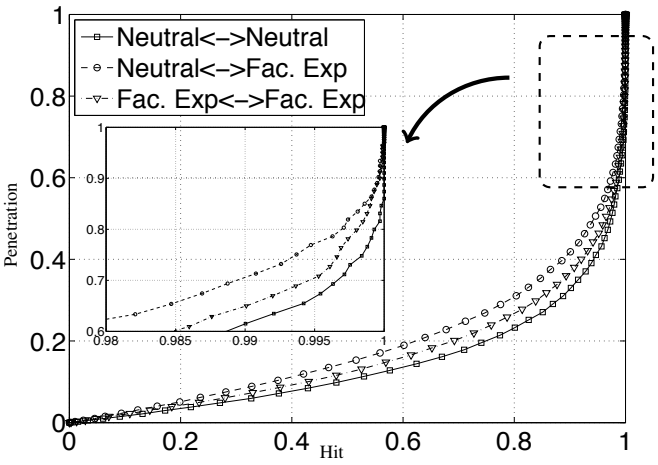


Fig. 20. Decreases in soft labels performance with respect to data of *non-neutral* facial expression. Using queries with *non-neutral* facial expression decreases the effectiveness of the soft labels, which can be counterbalanced if 3D head shapes with *non-neutral* expressions are also used (Fac. Exp. $\leftrightarrow$ Fac. Exp. series). Results are given for 15 $\times$ 15 SOMs.

Finally, the suitability of the proposed method recognizing facial expressions in multi-pose data was assessed. We doubled the number of 3D head hypotheses, having generated for each *neutral* head shape a corresponding *happy* expression (second row in Fig. 18). Joint head shapes / pose hypotheses were clustered and indexed in the same way as before. Next, for each SOM cell $\mathbf{s}_{c_i}$, the number of *neutral* / *happy* 3D head shape hypotheses associated with it was assumed to give the class likelihood $p(\mathbf{s}_{c_i}|\theta)$ in that region of the feature space, with $\theta \in \{$"Neutral","Happy"$\}$. Then, any query assigned to $\mathbf{s}_{c_i}$ was classified in terms of facial expression according to the Bayesian paradigm, with the posterior probability for a facial expression given by $p(\theta|\mathbf{s}_{c_i}) \propto p(\mathbf{s}_{c_i}|\theta)\ .\ p(\theta)/p(\mathbf{s}_{c_i})$. Under this formulation, and using equal priors per class, queries are considered to have *neutral* / *happy* expressions according to the most frequent expression of the 3D head shape hypotheses associated with that cell.

The left plot in Fig. 21 illustrates the power of cells in a 15 $\times$ 15 SOM to discriminate facial expressions, showing the $|\mathbf{s}_{c_i}^{(n)}|/|(\mathbf{s}_{c_i}^{(n)}| + |\mathbf{s}_{c_i}^{(h)}|)$ per cell, $|\mathbf{s}_{c_i}^{(.)}|$ being the number of 3D head shapes of *neutral* (n) / *happy* (h) expression associated with a cell. Values around 0.5 denote the non-interesting cases, i.e., cells with poor discriminating power (the number of *neutral* and *happy* elements is balanced). The right side of this same figure gives the confusion matrices for the LFW and CK+ sets, showing the mean and standard deviation performance values when repeating the recognition tests, using each time 85%

of the available samples in a bootstrapping-like strategy. The results are below the state-of-the-art [9] method, mostly due to the poor discriminating cells with classification performance only slightly better than random. In our view, results would be improved if facial models with more facial landmarks are used, which in poor quality data would be hard to infer without filtering techniques (e.g., graphical models to obtain the optimum configuration from a set of candidate landmarks). Note that filtering landmarks would violate one constraint in this paper: using exclusively *non-filtered* landmarks to enable real-time processing.

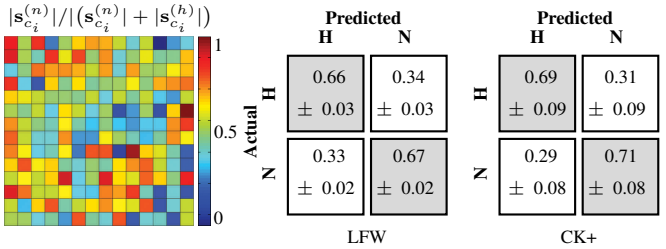


Fig. 21. At left: power of each SOM cell $\mathbf{s}_{c_i}$ to discriminate between *neutral* and *happy* expressions, expressed by the proportion of *neutral* head centroids associated to each cell. Red / orange cells represent predominantly *neutral* regions of the SOM space, whereas blue cells represent predominantly *happy* regions. Green / yellow cells have a balanced number of shapes per expression, making them particularly weak to discriminate between both classes. At right: confusion matrices for discriminating *neutral* / *happy* expressions in the LFW and CK+ sets.

VI. Conclusion

In this paper, we proposed a method to infer jointly human head poses and soft biometric labels based on the 3D morphology of the human head (the joint lengths between particular positions on the head). Using learning data from anthropometric surveys, a set of typical 3D head shapes (the labels) was inferred. Next, we described an algorithm to associate labels to low quality query samples, where subjects appear partially occluded and in varying poses. Using projective geometry techniques, we efficiently ranked a set of joint poses / head shape hypotheses, and iteratively evaluated the most likely hypothesis. The idea is to explicitly evaluate only a few hypotheses before the algorithm convergence, which is the key for the reduced temporal cost of the whole process.

The experiments were carried out using challenging data sets and support the usefulness of the soft biometric labels in two different ways: 1) coupled with a strong biometric classifier (e.g., a face recognizer), the resulting ensemble offers consistent improvements in performance over the strong expert alone; and, more importantly 2) these labels accord the concept of privacy-preserving recognition. In public environments, there are ethical / privacy issues behind the covert recognition of every subject passing-by. If soft labels are used, the system can confidently ignore the large majority of the identities in a scene and perform positive recognition only for a small subset of the subjects (those with soft labels similar to the watch-list elements).

References

[1] K. H. An and M. Chung. 3D head tracking and pose-robust 2D texture map-based face recognition using a simple ellipsoid model. In Proceedings of the *International Conference on Intelligent Robots Systems*, pag. 307–312, 2008.

[2] S. Arashloo and J. Kittler. Efficient processing of MRFs for unconstrained-pose face recognition. In Proceedings of the *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems*, pag. 1–8, 2013.

[3] S. Ba and J.-M. Odobez. Evaluation of multiple cue head pose estimation algorithms in natural environments. In Proceedings of the *IEEE International Conference on Multimedia and Expo*, pag. 1330–1333, 2005.

[4] R. Byrd, M. Hribar and J. Nocedal. An Interior Point Algorithm for Large-Scale Nonlinear Programming. *SIAM Journal on Optimization*, vol. 9, no. 4, pag. 877–900, 1999.

[5] E. Murphy-Chutorian and M. Trivedi. Head Pose estimation in Computer Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pag. 607–626, 2009.

[6] J. Dass, M. Sharma, E. Hassan and H. Ghosh. A Density Based Method for Automatic Hairstyle Discovery and Recognition. In Proceedings of the *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG'13)*, pag. 1–4, 2013.

[7] A. Drosou, D. Tzovaras, K. Moustakas and M. Petrou. Systematic Error Analysis for the Enhancement of Biometric Systems Using Soft Biometrics. *IEEE Signal Processing Letters*, vol. 19, no. 12, pag. 833–836, 2012.

[8] B. Efraty, C. Huang, S. Shah and I. Kakadiaris. Facial Landmark Detection in Uncontrolled Conditions. In Proceedings of the *2011 International Joint Conference on Biometrics*, pag. 1–8, 2011.

[9] S. Eleftheriadis, O. Rudovic and M. Pantic. Discriminative Shared Gaussian Processes for Multiview and View-Invariant Facial Expression Recognition. *IEEE Transactions on Image Processing*, vol. 24, no. 1, pag. 189– 204, 2015.

[10] M. Grgic, K. Delac and S. Grgic. SCface - surveillance cameras face database. *Multimedia Tools and Applications Journal*, vol. 51, no. 3, pag. 863–879, 2011.

[11] D. Heckathorn, R. Broadhead and S. Sergeyev. Anthropometry of Flying Personnel-1950. Technical report No. 52-321, Wright Air Development Center, Wright Patterson Air Force Base, Ohio, 1954.

[12] H. Hertzberg, G. Daniels and E. Churchill. A methodology for reducing respondent duplication and impersonation in samples of hidden populations. In Proceedings of the *Annual Meeting American Sociology Association*, pag. 543–564, 1997.

[13] J. Hewig, R. Trippe, H. Hecht, T. Straube and W. Miltner. Gender differences for specific body regions when looking at men and women. *Journal of Nonverbal Behaviour*, vol. 32, no. 2, pag. 67–78, 2008.

[14] K. Huang and M. Trivedi. Robust real-time detection, tracking and pose estimation of faces in video streams. In Proceedings of the *International Conference on Pattern Recognition*, pag. 965-968, 2004.

[15] G. Huang, V. Jain and E. Learned-Miller. Unsupervised joint alignment of complex images. In Proceedings of the *International Conference on Computer Vision*, pag. 1-8, 2007.

[16] A. K. Jain, S. C. Dass and K. Nandakumar. Can soft biometric traits assist user recognition? In Proceedings of the *SPIE*, vol. 5404, pag. 561-572, 2004.

[17] A. K. Jain and U. Park. Facial marks: Soft biometric for face recognition. In Proceedings of the *IEEE International Conference on Image Processing* (ICIP'09), pag. 37-40, 2009.

[18] S. Jaiswal, T. Almaev and M. Valstar. Guided Unsupervised Learning of Mode Specific Models for Facial Point detection in the Wild. In Proceedings of the *IEEE International Conference on Computer Vision Workshops*, pag. 370–377, 2013.

[19] M. Koestinger, P. Wohlhart, P. Roth and H. Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In Proceedings of the *First IEEE International Conference on Computer Vision Workshops*, pag. 2144–2151, 2011.

[20] T. Kohonen. Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, vol. 43, no. 1, pag. 59–69, 1982.

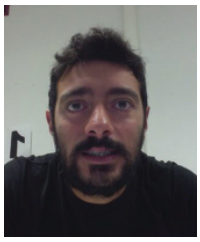
[21] J. Lagarias, J. Reeds, M. Wright and P. Wright. Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *SIAM Journal of Optimization*, vol. 9, no. 1, pag. 112–147, 1998.

[22] Y. Li, S. Gong, J. Sherrah and H. Liddell. Support vector machine based multi-view face detection and recognition. *Image and Vision Computing*, vol. 1, no. 5, pag. 413–427, 2004.

[23] T. Lucas and M. Henneberg. Comparing the face to the body, which is better for identification? *International Journal of Legal Medicine*, doi: 10.1007/s00414-015-1158-6, 2015.
[24] P. Lucey, J. Cohn, T. Kanade, J. Saragih and Z. Ambadar. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression In Proceedings of the *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pag. 94–101, 2010.
[25] M-G. Kim, H-M. Moon, Y. Chung and S. Pan. A Survey and Proposed Framework on the Soft Biometrics Technique for Human Identification in Intelligent Video Surveillance System. *Journal of Biomedicine and Biotechnology*, doi: 10.1155/2012/614146, 2012.
[26] M. Krinidis, N. Nikolaidis and I. Pitas. 3-D Head Pose Estimation in Monocular Video Sequences Using Deformable Surfaces and Radial Basis Functions. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, issue 2, pag. 161–272, 2009.
[27] N. Kruger, M. Potzsch and C. von der Malsburg. Determination of Face Position and Pose with a Learned Representation Based on Labeled Graphs. *Image and Vision Computing*, vol. 15, no. 8, pag. 665–673, 1997.
[28] B. Kwolek. Model based facial pose tracking using a particle filter. In Proceedings of the *Geometric Modelling and Imaging - New Trends Conference*, pag. 203–208, 2006.
[29] K. Moustakas, D. Tzovaras and G. Stavropoulos. Gait Recognition Using Geometric Features and Soft Biometrics. *IEEE Signal Processing Letters*, vol. 17, no. 4, pag. 367–370, 2010.
[30] F. Moreno-Noguer, V. Lepetit and P. Fua. Pose priors for simultaneously solving alignment and correspondence. In Proceedings of the *European Conference on Computer Vision*, part II, pag. 405–418, 2008.
[31] R. Osadchy, M. Miller and Y. LeCun. Synergistic Face Detection and Pose Estimation with Energy-Based Models. *Journal of Machine Learning Research*, vol. 8, pag. 1197–1215, 2007.
[32] U. Park and A. K. Jain. Face matching and retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pag. 406–415, 2010.
[33] V. Rapp, T. Senechal, K. Bailly and L. Prevost. Multiple kernel learning SVM and statistical validation for facial landmark detection. In Proceedings of the *2011 IEEE International Conference on Automatic Face and Gesture Recognition*, pag. 265-271, 2011.
[34] D. Reid, S. Samangooei, C. Chen, M. Nixon and A. Ross. Soft Biometrics for Surveillance: An Overview. Handbook of Statistics, vol. 31, pag. 327-351, 2013.
[35] D.A. Reid, M.S. Nixon and S. Stevenage. Soft Biometrics; Human Identification Using Comparative Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pag. 1216–1228, 2014.
[36] A. Rice, P.J. Phillips and A. O'Toole. The Role of the Face and Body in Unfamiliar Person Identification. *Applied Cognitive Psychology*, vol. 27, issue 6, pag. 761–768, 2013.
[37] J. Sánchez-Riera, J. Öslund, P. Fua and F. Moreno-Noguer. Simultaneous Pose, Correspondence and Non-Rigid Shape. In Proceedings of the *2010 IEEE Conference on Computer Vision and Pattern Recognition*, pag. 1189–1196, 2010.
[38] J. Sherrah, S. Gong and E.-J. Ong. Face Distributions in Similarity Space under Varying Head Pose. *Image and Vision Computing*, vol. 19, no. 12, pag. 807–819, 2001.
[39] E. Tola, V. Lepetit and P. Fua. Daisy: An efficient dense descriptor applied to wide baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pag. 815–830, 2010.
[40] P. Tome, J. Fierrez, R. Vera-Rodriguez and M. Nixon. Soft Biometrics and Their Application in Person Recognition at a Distance. *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pag. 464–475, 2014.
[41] J. Tu, T. Huang and H. Tao. Accurate head pose tracking in low resolution video. In Proceedings of the *International Conference on Automatic Face and Gesture Recognition*, pag. 573–578, 2006.
[42] J. Xiao, S. Baker, I. Matthews and T. Kanade. Real-Time Combined 2D / 3D Active Appearance Models. In Proceedings of the *International Conference on Computer Vision and Pattern Recognition*, vol. 2, pag. 535–542, 2004.
[43] J. Wu and M. Trivedi. A Two-Stage Head Pose Estimation Framework and Evaluation. *Pattern Recognition*, vol. 41, no. 3, pag. 1138–1158, 2008.
[44] J.W. Young. Head and Face Anthropometry of Adult U.S. Civilians. Office of Aviation medicine, Federal Aviation Administration, DOT/FAA/AM-93/10, 1993.
[45] C. Zhang and F. Cohen. 3-d face structure extraction and recognition from images using 3-d morphing and distance mapping. *IEEE Transactions on Image Processing*, vol. 11, pag. 1249–1259, 2002.
[46] X. Zhang and Y. Gao. Face recognition across pose: A review. *Pattern Recognition*, vol. 42, issue 11, pag. 2876–2896, 2009.
[47] Z. Zhu and Q. Ji. Robust Real-Time Face Pose and Facial Expression Recovery. In Proceedings of the 2006 IEEE Computer Society on Computer Vision and Pattern Recognition, vol. 1, pag. 681–688, 2006.
[48] X. Zhu and D. Ramanan. Face Detection, Pose Estimation, and Landmark Localization in the Wild. In Proceedings of the 2012 IEEE Computer Society on Computer Vision and Pattern Recognition, pag. 2879–2886, 2012.

Hugo Proença B.Sc. (2001), M.Sc. (2004) and Ph.D. (2007) is an Associate Professor at University of Beira Interior and has been researching mainly about biometrics and visual-surveillance. He is the coordinating editor of the IEEE Biometrics Council Newsletter and the area editor (ocular biometrics) of the IEEE Biometrics Compendium Journal. He is a member of the Editorial Board of the International Journal of Biometrics and served as Guest Editor of special issues of the Pattern Recognition Letters, Image and Vision Computing and Signal, Image and Video Processing journals.

João C. Neves received the B.Sc. and M.Sc. degrees in Computer Science from the University of Beira Interior, Portugal, in 2011 and 2013, respectively. He is currently working towards the Ph.D. degree from the same university in the area of biometrics. His research interests include computer vision and pattern recognition, with a particular focus on biometrics and surveillance.



Silvio Barra was born in 1985 in Battipaglia (Salerno, ITALY). In 2009 and in 2012 he received the B.Sc. degree (cum laude) and the M.Sc. degree (cum laude) in Computer Science from University of Salerno. Since December 2012 he is a Ph.D. Student at the University of Cagliari. His main research interests include pattern recognition, biometrics and video analysis and analytics.

Tiago Marques has three year frequency in Management at Instituto de Economia e Gestão, Lisbon, Portugal between 2008 and 2011. Worked at *Induscria*, a Creative Industry's cross-platform association where he coordinated projects between multiple companies, in 2012. He is currently an undergraduate student at the Computer Science at Universidade da Beira Interior and a researcher at the SOCIA Lab.

Juan C. Moreno is a research fellow at the Pattern and Image Analysis Group of the University of Beira Interior, Portugal. He received his B.Sc. and M.Sc Degrees in Mathematics from the Central University of Venezuela (2004) and from Simón Bolívar University (2008), respectively. In 2013, he completed his Ph.D. in Mathematics at the University of Coimbra. His interests revolve around mathematical based methods for image processing, computer vision and pattern recognition problems.