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/TIFS.2017.2680246, IEEE Transactions on Information Forensics and Security

Soft Biometrics: Globally Coherent Solutions for Hair Segmentation and Style Recognition based on Hierarchical MRFs

Hugo Proença, *Senior Member, IEEE* and João C. Neves, *Member, IEEE*

***Abstract*—Markov Random Fields (MRFs) are a popular tool in many computer vision problems and faithfully model a broad range of local dependencies. However, rooted in the Hammersley-Clifford theorem, they face serious difficulties in enforcing the global coherence of the solutions without using too high order cliques that reduce the computational effectiveness of the inference phase. Having this problem in mind, we describe a multi-layered (hierarchical) architecture for MRFs that is based exclusively in pairwise connections and typically produces globally coherent solutions, with 1) one layer working at the local (*pixel*) level, modelling the interactions between adjacent image patches; and 2) a complementary layer working at the *object* (hypothesis) level pushing toward globally consistent solutions. During optimization, both layers interact into an equilibrium state, that not only segments the data, but also classifies it. The proposed MRF architecture is particularly suitable for problems that deal with biological data (e.g., biometrics), where the reasonability of the solutions can be objectively measured. As test case, we considered the problem of hair / facial hair segmentation and labelling, which are soft biometric labels useful for human recognition *in-the-wild*. We observed performance levels close to the state-of-the-art at a much lower computational cost, both in the segmentation and classification (labelling) tasks.**

***Index Terms*—Soft Biometrics, Visual Surveillance, Homeland Security.**

I. Introduction

IN visual surveillance / biometrics research, the development of systems to work in unconstrained data acquisition protocols and uncontrolled lighting environments is a major ambition. The images resulting of such conditions are degraded in multiple ways, such as blurred, shadowed, of poor resolution, with subjects off-angle and partially occluded (Fig. 1). In these cases, soft biometrics can be seen as an identity retrieval tool that attenuates the decrease in performance of the classical biometric traits (e.g., the face or the iris).

The descriptions of the facial hair and hair styles are among the most effective soft biometric traits reported in the literature [24]. In this scope, the pioneer analysis methods were designed to work exclusively in good quality images of frontal subjects. Regardless recents attempts to increase the robustness (e.g., [29]), the ambition of working effectively in images acquired in typical visual surveillance conditions remains to be achieved.

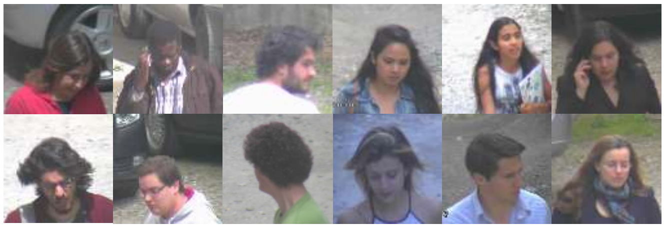

Fig. 1. Examples of images captured by an outdoor visual surveillance system, with *unconstrained* acquisition conditions and protocols. Images have typically poor resolution and are often blurred, with subjects partially occluded and under varying poses.

Markov Random Fields (MRFs) are a classical tool for many computer vision problems, from image segmentation [13], image registration [8] to object recognition [4]. Among other strengths, they provide non-causal models with isotropic behaviour and faithfully model a broad range of local dependencies. On the other way, they hardly guarantee globally coherent solutions without using too high order cliques that compromise the computational effectiveness of the inference phase. Having this problem in mind, in this paper we propose a multi-layered (hierarchical) MRF that does not use high order cliques but still typically reaches globally coherent solutions. As test case, we consider the hair / facial hair style analysis, and describe an inference process composed of two phases:

1) three supervised non-linear classifiers run at the pixel level and provide the posterior probabilities for each image position and class of interest: *hair*, *skin* and *background*. Each classifier detects one component based on texture and shape image statistics;
2) the posteriors based on data *appearance* are combined with geometric constraints and a set of model hypotheses to feed the MRF, composed of a *segmentation* and a *classification* layer. One layer discriminates locally the classes of interest, while the other infers the soft biometric labels that describe the query's facial hair and hair styles.

The key idea is to combine the strengths of MRFs with groups of synthetic hypotheses that are projected onto the input plane and guarantee the global consistency (biological coherence) of the solution. The proposed model inherits some insights from previous works that used shape priors to constraint the models (e.g., [2]) and multiple layered MRFs (e.g., [26] and [20]).

Authors are with the IT: Instituto de Telecomunicações, Department of Computer Science, University of Beira Interior, Covilhã, Portugal, E-mail: {hugomcp, jcneves}@di.ubi.pt. This work was supported by FCT project UID/EEA/50008/2013.

Manuscript received ? ?, 2016; revised ? ?, ?.

1556-6013 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TIFS.2017.2680246, IEEE Transactions on Information Forensics and Security

The remainder of this paper is organized as follows: Section II analyzes the related work. Section III details the learning and inference phases of the proposed method. Section V describes our experiments and the conclusions are given in Section VI.

II. Related Work

Table I overviews the literature for facial hair / hair style analysis. Algorithms are classified according to their scope (Hair (H) / Facial Hair (FH), Segmentation / Classification), along with a description of the techniques / color spaces used. The data variability factors considered are enumerated, with Y, P and R denoting deviations in *yaw*, *pitch* and *roll* angles, and A and C referring the abilities to work with unaligned data and unconstrained hair colors. Below, methods are grouped into three families: predominantly generative, discriminative and hybrid.

Lee *et* al. [17] propose a generative model that infers a set of hypotheses for the face, hair, and background regions. In classification, the most reliable pixels are the information source for mixture models that parameterise each component and define the MRF unary costs. Still in the generative family, Shen and Ai [23] propose a face detector to define the ROI and consider color information (YCbCr space) to feed a MRF used for segmentation. As post-processing, nearest neighbour analysis enforces the homogeneity between adjacent regions. Wang *et* al. [28] formulate the segmentation problem as finding pairs of isomorphic manifolds, using a set of learning images with the corresponding ground-truth, designated as *optimal maps*. Here, queries are represented as combinations of optimal maps. In [33] [34] Zhang *et* al. infer a set of probability density functions of four typical hair colors (XYZ and HSV spaces), learned by the expectation-maximization algorithm. Assuming the statistical independence between color channels, they obtain the likelihood in each color space and use a Bayesian framework to segment hair. Finally, a simple approach is due to Dass *et* al. [5], that segment the hair regions by thresholding and use agglomerative clustering to parameterise five groups of hairstyles, based on the proportion of hair pixels in image patches.

Regarding methods that are predominantly discriminative, Kae *et* al. [12] detect the most homogenous image patches (*super-pixels*), which provide the appearance information to a CRF. To guarantee the global coherence of the hypotheses, a restricted Boltzmann machine encodes the global shape priors and enforces shape constraints. Wang and Ai [27] learn a discriminator between the hair / non-hair regions. In classification, seven hairstyles are considered, with the RankBoost algorithm selecting the most informative patches and defining hairstyle similarity directly on the hair shapes. Under the same paradigm, Rouset and Coulon [22] fuse color (YCbCr space) to frequency information, in order to locally discriminate between hair / non-hair pixels.

Hybrid approaches are typically based in template matching, with the pioneer method due to Yacoob and Davis [30]. These authors use face and eye detectors to define the ROIs. Based on spatial and color information, a set of seeds is inferred and region growing is used based on local homogeneity. Finally, morphologic operators enforce connected components. Julian *et* al. [11] learn a set of shape templates of the upper part of the head, based on the boundary control points. Using principal components analysis, they propose the concept of *eigen shape*, keeping the top variability vectors that represent the 3D head orientation and the face morphology. Hair regions are classified at the pixel level according to a texture-analysis strategy, generating seeds for subsequent finer parameterisations (active contours). Ugurlu [25] use a head pose detector based both in shape and texture, being the latter described statistically in the HSV color space. Wang *et* al. [29] use a head detector that defines a ROI, based in histogram analysis and nearest neighbour rules. The hair length is inferred by line scanning on the segmented hair region. A relevant gap of this work is the fact of being only suitable for handling dark hair subjects. Lipowezky *et* al. [18] start by detecting head landmarks (eyes and mouth) to find the most homogenous image patches. Color information (LAB and YCbCr spaces) is fused to the Canny magnitude and to four texture descriptors (wavelets-based), feeding a region-growing algorithm. Similarly, Krupka *et* al. [16] use a head detector that defines the ROI where the skin is detected and segmented. The differences between the head foreground and the skin pixels provide the estimate of the hair positions. Skin seeds are detected by thresholding, further expanded upon homogeneity.

III. Proposed Method

For comprehensibility, the following notation is adopted: matrices are represented by capitalized bold font and vectors appear in bold. The subscripts denote indexes. All vectors are column-wise. The ring symbol (e.g., $\mathring{\mathbf{x}}$) denotes image positions, while 3D positions appear in regular font (e.g., $\mathbf{x}$).

A. Synthesis of 3D Models

We consider three types of 3D models: 1) head; 2) hair; and 3) facial hair. The head models are generated as described in [19]. Using the Young's [31] head anthropometric survey to obtain a group of probability density functions of human head lengths and a basis 3D mesh, we deform the mesh according to randomly drew target distances between pairs of vertices (l_{ij}). Let $\mathbf{x}_i$ be one 3D vertex and $\mathbf{n_i}$ the normal to the surface at that point. Let $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$, $\mathbf{n}_{ij} = \mathbf{n}_i - \mathbf{n}_j$ ($\mathbf{x}_., \mathbf{n}_. \in \mathbb{R}^3$) and let l_{ij} be the target length (Euclidean distance) between $\mathbf{x}_i$ and $\mathbf{x}_j$. This yields a system of linear equations with inequality constraints, enabling to find (using [3]) the magnitude of the displacement α_{ij} on both vertices with respect to their normals ($\mathbf{x}^{\text{new}} = \mathbf{x}^{\text{old}} + \alpha\mathbf{n}$), such that their distance is l_{ij} and $||\boldsymbol{\alpha}||_\infty \leq \kappa_0$ (to avoid anatomically bizarre solutions). The top row of Fig. 2 illustrates our population of head shapes.

Let $\mathbf{s}_s = [\mathbf{x}_1^T, \ldots, \mathbf{x}_{t_v}^T]^T$ be a vector representing one head shape, given as a triangulated mesh of t_v vertices. Considering a set of head shapes $\mathbf{S}_s = \{\mathbf{s}_{s,1}, \ldots, \mathbf{s}_{s,t_m}\}$, there is evidently strong correlation between the $\mathbf{x}_i$ elements in those meshes, which is attenuated if they are represented in the principal components (PC) space:

1556-6013 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TIFS.2017.2680246, IEEE Transactions on Information Forensics and Security

TABLE I
SUMMARY OF THE MOST RELEVANT METHODS TO PERFORM AUTOMATED DETECTION, SEGMENTATION AND CLASSIFICATION OF FACIAL HAIR AND HAIR STYLES. Y, P, R , A AND C STAND FOR THE DATA VARIATION FACTORS EACH METHOD CLAIMS TO HANDLE: DEVIATIONS IN YAW, PITCH AND ROLL ANGLES, UNALIGNED DATA AND NON-EXISTENCE OF HAIR COLOR CONSTRAINTS.

Method	**Year**	**Type**	**Working Mode**		**Data Variability**					**Color Sp.**	**Summary**
			Segm.	**Class.**	**Y**	**P**	**R**	**A**	**C**		
Yacoob and Davis [30]	2006	H	✓	✓	✗	✗	✗	✗	✓	RGB	Gabor kernels (hair texture), dominant color, anthropometric statistics
Lee *et* al. [17]	2008	H	✓	✗	✗	✗	✗	✗	✓	RGB, LAB	Graphical model
Lipowezky *et* al. [18]	2008	H	✓	✗	✗	✗	✗	✗	✓	LAB, HSV	Seed detection: EDISON algorithm, edge analysis, Haar filtering, region growing: k-means clustering
Rouset and Coulon [22]	2008	H	✓	✗	✗	✗	✗	✗	✓	HSV, YCbCr	Frequency analysis (Gaussian filtering), color analysis (local deviations)
Zhang *et* al. [33]	2008	H	✓	✗	✓	✓	✓	✗	✓	HSV, XYZ	Gaussian mixture model-based density estimation
Zhang *et* al. [34]	2009	H	✓	✗	✓	✓	✓	✓	✓	HSV, XYZ	Gaussian mixture density estimation, analysis of skin, hair and head spatial constraints
Julian *et* al. [11]	2010	H	✓	✗	✗	✗	✗	✗	✓	HSV, YCbCr	Histogram analysis, active shape models
Wang and Ai [27]	2013	H	✓	✓	✗	✗	✓	✗	✓	LUV	Informative patches (RankBoost, SVM), graphical model, clustering
Ugurlu [25]	2012	H, FH	✓	✗	✓	✗	✗	✗	✓	HSV	Non-linear supervised local classification
Dass *et* al. [5]	2013	H	✓	✓	✓	✗	✗	✓	✓	HSV, grayscale	Eyes detection (Adaboost), alignment (similarity transform), Otsu thresholding, clustering
Kae *et* al. [12]	2013	H, FH	✓	✓	✓	✓	✓	✗	✓	RGB	Conditional random fields (local consistency), restricted Boltzmann machines (global consistency)
Wang *et* al. [28]	2013	H	✓	✗	✓	✓	✓	✓	✓	RGB	Coarse local likelihood, manifold inference, refined segmentation
Krupka *et* al. [16]	2014	H	✓	✗	✗	✗	✗	✗	✓	RGB	Background modelling (Gaussian mixture), convex hull analysis
Shen and Ai [23]	2014	H	✓	✓	✗	✗	✗	✗	✓	YCbCr	Landmarks detection (ASM), graph-cuts, histogram analysis
Wang *et* al. [29]	2014	H	✓	✓	✓	✓	✓	✓	✗	Grayscale	Thresholding, histogram analysis, line intersection)
Proposed Method	2016	H, FH	✓	✓	✓	✓	✓	✓	✓	RGB, HSV, YCbCr	Non-linear pixel classification, 3D Model projection, hierarchical graphical model, manifold learning

$$\mathbf{s}_s^* = (\mathbf{s}_s - \bar{\mathbf{s}_s})\mathbf{T}_{pc}, \tag{1}$$

where $\bar{\mathbf{s}_s}$ is the $3t_v$-dimensional mean of the elements in $\mathbf{S}_s$ and $\mathbf{T}_{pc}$ is the PC transformation matrix. This way, each mesh is represented in a feature space of a much lower dimension than the $3t_v$, which accounts for the computational effectiveness of the whole method. In our case, the head models have $t_v = 957$, but 50 PC coefficients represent over 99.9% of the variability.

Regarding the hair / facial hair models, we use the concept of *hair mesh* from Yuksel *et* al. [32] and consider hair / facial hair classes as particular cases of polygonal mesh modelling. For simplicity, we keep a short number of hypotheses for each class: $\mathbf{S}_h$={”*bald*”, ”*short bald*”, ”*short*”, ”*medium*”, ”*long fine*”, ”*long volume*”} for the hair and $\mathbf{S}_f$={”*clean*”, ”*moustache*”, ”*goatee*”, ”*beard*”} for the facial hair. As previously, all models are generated by deforming iteratively a basis 3D mesh (examples are shown at the bottom rows of Fig. 2).

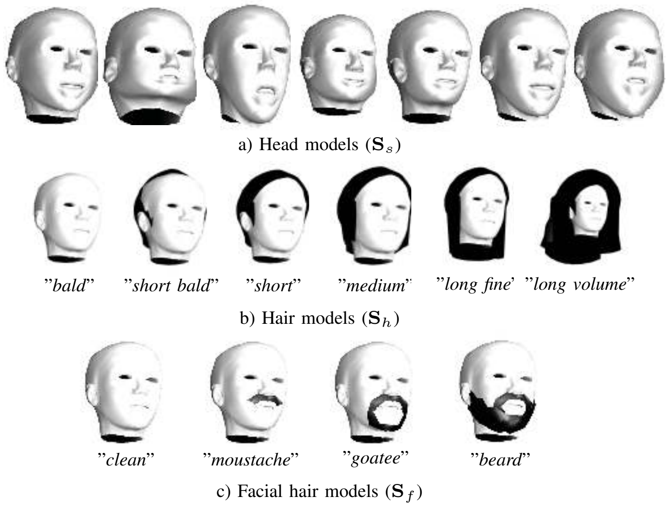
a) Head models ($\mathbf{S}_s$)
b) Hair models ($\mathbf{S}_h$)
c) Facial hair models ($\mathbf{S}_f$)

Fig. 2. Illustration of the 3D head shapes, hair and facial hair models that are used as the hypotheses considered in this paper.

B. Pose Hypotheses

We also consider a set of 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 along the orthogonal axes t_x, t_y and t_z. $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_{t_p}\}$ is the set of t_p pose hypotheses uniformly distributed over all the six degrees of freedom.

C. Joint Head Shape / Pose Hypotheses Indexing

Given a set of t_p pose and t_s head shape hypotheses, it is required to find the best joint pose / head shape configuration, which will most likely match the query. To avoid exploring by brute-force all $t_p t_s$ possibilities, a forest of binary trees

1556-6013 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TIFS.2017.2680246, IEEE Transactions on Information Forensics and Security

is created at learning time, one tree per type of landmark. In these indexing structures, hypotheses are grouped (k-means) in branches according to the neighbourhood of the projected landmark. The *world-to-image* function projects the $\mathbf{x}$ vertices of a head shape hypothesis $\mathbf{s_s}$ according to a pose configuration $\mathbf{p}$:

$$f_{w \to i}(\mathbf{x}, \mathbf{p}) = \mathring{\mathbf{x}} = \frac{1}{\upsilon}\mathbf{A}[\mathbf{R}|\mathbf{t}]\begin{bmatrix}\mathbf{x}\\1\end{bmatrix}, \tag{2}$$

where υ is the scalar projective parameter, $\mathbf{A}$ is the internal camera matrix, and $\mathbf{R}$ and $\mathbf{t}$ are the pose parameters. The retrieval time of the forest is approximately logarithmic with respect to the number of hypotheses, which enables to generate a large set of hypotheses without compromising the time cost of retrieval. Additional details about this data structure are given in [19].

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 part 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., [10], [6], or [21]). The position of every query landmark enters in the corresponding binary tree to retrieve the indices of the complying hypotheses. By accumulating the complying indices over all trees, the hypotheses are ranked in descending order according to the likeliness they match the query. Refer to [19] for full details about the way the most likely head shape $\hat{\mathbf{s}}$ and pose $\hat{\mathbf{p}}$ hypotheses are inferred. Fig. 3 gives examples of the head shape / pose estimation inference, using images of the AFLW [15] set. The five leftmost columns contain successful cases, whereas the rightmost column illustrates failure cases, mostly due to ambiguities in various head shape / pose configurations that provide too many overlapped landmark projections.

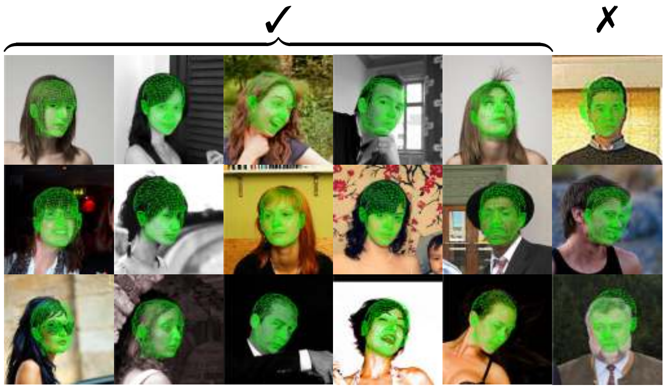

Fig. 3. Successful / failure (rightmost column) estimates of the head shape / pose. Most failed cases are due to ambiguities in various head shape / pose configurations that provide too many overlapped landmark projections.

IV. Soft Labels Inference

After inferring the query head shape $\hat{\mathbf{s}}$ and pose $\hat{\mathbf{p}}$ hypotheses, all hair / facial hair hypotheses are projected according to $\{\hat{\mathbf{s}}, \hat{\mathbf{p}}\}$, to perceive how much they agree with the data appearance terms, which implicitly constitute part of the MRF costs. Fig. 4 gives a cohesive perspective of the two-layered MRF we propose. One layer works at the pixel level (segmentation layer), with a bijection between image pixels and nodes, each one with three potential labels: *hair*, *skin* and *background*. The other layer (classification) has two nodes that represent the facial hair / hair hypotheses. During model optimization, the interaction between both layers privilege pixel labels that accord a parameterization of the classification nodes and *vice-versa*, forcing the network to converge into an equilibrium state where the configurations at one layer implicitly segment data and the parameterizations in the other layer enforce biologic coherent solutions and describe the facial hair / hair styles.

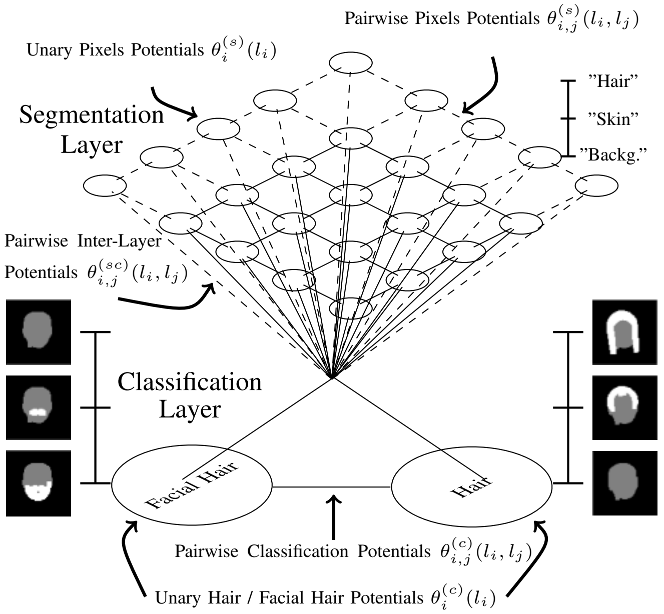
Pairwise Pixels Potentials $\theta^{(s)}_{i,j}(l_i, l_j)$
Unary Pixels Potentials $\theta^{(s)}_{i}(l_i)$
Segmentation Layer
”Hair”
”Skin”
”Backg.”
Pairwise Inter-Layer Potentials $\theta^{(sc)}_{i,j}(l_i, l_j)$
Classification Layer
Facial Hair
Hair
Pairwise Classification Potentials $\theta^{(c)}_{i,j}(l_i, l_j)$
Unary Hair / Facial Hair Potentials $\theta^{(c)}_{i}(l_i)$

Fig. 4. Structure of the MRF that fuses the data appearance information (upper layer) to global constraints (bottom layer). During optimization, the the network should converge into a balance point where the predominant labels at the segmentation level are biologically plausible and accord globally coherent facial hair / hair hypotheses (at the classification level).

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph representing a MRF of t_v vertices $\mathcal{V}$, linked by t_e edges $\mathcal{E}$. Let t_s be the number of vertices in the *segmentation* layer and t_c the number of vertices in the *classification* layer, such that $t_v = t_s + t_c$. The MRF is a representation of a discrete latent random variable $\boldsymbol{L} = \{L_i\}, \forall i \in \mathcal{V}$, where each element L_i takes one value l_i from a set of labels. Let $\boldsymbol{l} = \{l_1, \ldots, l_{t_s}, l_{t_s+1}, \ldots, l_{t_s+t_c}\}$ represent one configuration of the MRF. In our model, the classification nodes are connected to each other and to all pixel nodes, while the pixel nodes are connected to their horizontal / vertical neighbours. Note that the proposed model does not use high-order cliques. Even though there is a point in Fig. 4 that joins multiple edges, it actually represents overlapped pairwise connections between one classification and one segmentation node.

The energy of a configuration $\boldsymbol{l}$ of the MRF is the sum of

1556-6013 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TIFS.2017.2680246, IEEE Transactions on Information Forensics and Security

the unary $\theta_i(l_i)$ and pairwise $\theta_{i,j}(l_i, l_j)$ potentials:

$$E(\boldsymbol{l}) = \sum_{i \in \mathcal{V}} \theta_i(l_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{i,j}(l_i, l_j). \tag{3}$$

According to this formulation, segmenting / classifying an image is done by inferring the random variables that minimize its energy:

$$\hat{\boldsymbol{l}} = \arg\min_{\boldsymbol{l}} E(\boldsymbol{l}), \tag{4}$$

where $\{\hat{l}_1, \dots, \hat{l}_{t_s}\}$ are the labels of the pixels and $\{\hat{l}_{t_s+1}, \dots, \hat{l}_{t_s+t_c}\}$ specify the parameterizations in the classification nodes.

A. Feature Extraction

The data appearance is analyzed at the pixel level, to distinguish between three components in the image: hair, skin and background (any remaining information). As the red / blue chroma values provide good separability between skin and non-skin pixels [1] and the hair is frequently discriminated by analysing the HSV / RGB triplets (Table I), we extract, for each image pixel, a feature set composed of 81 elements: {red, green and blue channels (RGB); hue, saturation and value channels (HSV); red and blue chroma (yCbCr); LBP from the value channel}, considering the average, standard deviation and range statistics in square patches of side {5, 9, 15} around the central element.

B. Learning

1) Unary Potentials: Let $\gamma : \mathbb{N}^2 \rightarrow \mathbb{R}^{81}$ be the feature extraction function that produces a vector $\gamma(x, y) \in \mathbb{R}^{81}$ for each pixel at position (x, y). Let $\Gamma = [\gamma(x_1, y_1), \dots, \gamma(x_n, y_n)]^T$ be a $n \times 81$ matrix in a learning set used to create three non-linear binary classification models, one for each component $\omega_i \in \{$"Hair", "Skin", "Background"$\}$. Let $\eta_i : \mathbb{R}^{81} \rightarrow [0, 1]$ be the response of the ith model, regarded as an estimate of the class likelihood $P\Big(\eta_i\big(\gamma(x, y)\big)|\omega_i\Big)$. According to the Bayes rule, and assuming equal priors, the posterior probabilities are given by:

$$P\Big(\omega_i|\eta_i\big(\gamma(x, y)\big)\Big) = \frac{P\Big(\eta_i\big(\gamma(x, y)\big)|\omega_i\Big)}{\sum_{j=1}^{3} P\Big(\eta_j\big(\gamma(x, y)\big)|\omega_j\Big)}. \tag{5}$$

In our model, the unary potentials of the vertices in the segmentation layer are defined as $\theta_i^{(s)}(l_i) = 1 - P\Big(\omega_i|\eta_i\big(\gamma(x, y)\big)\Big)$. The unary potentials in the classification layer correspond to the agreement (exclusive-or) between the index of the maximum posterior probability at each point $I_m(x, y) = \arg\max_j p\Big(\omega_j|\eta_j\big(\gamma(x, y)\big)\Big)$ and the 3D model projections $I_p(x, y)$ obtained by the *world-to-image* function (2):

$$\theta_i^{(c)}(l_i) = \frac{1}{h\ w} \sum_{y=1}^{h} \sum_{x=1}^{w} \big(1 - \delta(I_m(x, y), I_p(x, y))\big), \tag{6}$$

with $\delta(.,.)$ being the Kronecker delta function, h and w the query height and width. The rationale here is to privilege the hair and facial hair models that provide the maximum overlap between the responses of the non-linear models and the projections of the corresponding 3D meshes.

2) Pairwise Potentials: There are three types of pairwise potentials in our model: 1) between segmentation nodes; 2) between classification nodes; and 3) between inter-layer nodes. The pairwise potentials between segmentation nodes $\theta_{i,j}^{(s)}(l_i, l_j)$ correspond to the prior probability of observing labels l_i, l_j in adjacent positions of a learning set, to privilege smooth solutions:

$$\theta_{i,j}^{(s)}(l_i, l_j) = \frac{1}{\kappa_1 + P(\mathfrak{C}(x', y') = \omega_i, \mathfrak{C}(x, y) = \omega_j)}, \tag{7}$$

where $P(.,.)$ is the joint probability, (x', y') and (x, y) are 4-adjacent positions, $\mathfrak{C}(.,.)$ denotes the component label {hair, skin, background} at one position and $\kappa_1 \in \mathbb{R}^+$ avoids infinite costs.

The pairwise potentials between classification nodes $\theta_{i,j}^{(c)}(l_i, l_j)$ consider the prior probabilities of observing two facial hair and hair hypotheses in the learning set (e.g., beards are more probable in bald and short hair than in long / long volume subjects).

The pairwise potentials between inter-layer nodes $\theta_{i,j}^{(sc)}(l_i, l_j)$ enforce the biological plausibility of the solution (8) and privilege the consistency between the configurations in both layers. This is done by penalising parameterisations of pixel nodes that are outside of the polygons defined by the boundaries of the projections of the head, hair and facial hair models (e.g., it is too costly to observe a *hair* pixel and a *bald* hypothesis).

$$\theta_{i,j}^{(sc)}(l_i, l_j) = \begin{cases} 0, \text{ if } \delta\Big(\delta(l_i, \mathfrak{C}'(i, l_j)), \\ \qquad\qquad \delta(\psi_t(x_i, y_i, \mathbf{x}_j, \mathbf{y}_j), 0\Big) = 0 \\ \text{erf}\Big(\kappa_2\ \psi_d(x_i, y_i, \mathbf{x}_j, \mathbf{y}_j)\Big), \text{ otherwise} \end{cases},$$

where $\psi_t(x_i, y_i, \mathbf{x}_j, \mathbf{y}_j) : \mathbb{N}^2 \times \mathbb{N}^n \rightarrow \{0, 1\}$ is an indicator function that assumes a unit value when the point (x_i, y_i) is inside the polygon defined by vertices $\{\mathbf{x}_j, \mathbf{y}_j\} = \{(x_{j,k}, y_{j,k})\}$. $\psi_d(x_i, y_i, \mathbf{x}_j, \mathbf{y}_j) : \mathbb{N}^2 \times \mathbb{N}^n \rightarrow \mathbb{R}^+$ is the point-to-polygon distance divided by the image diagonal length. erf$(.) : \mathbb{R}^+ \rightarrow [0, 1]$ is a transfer function (error function) with sigmoid shape with κ_2 controlling its shape (larger values lead to farther from linear shapes). Here, $\mathfrak{C}'(i, l_j)$ denotes the component label (hair, skin or background) at the i^{th} image position under the j^{th} joint facial hair / hair hypothesis. Fig. 5 illustrates the rationale of this kind of costs: for two queries, the responses given by the three non-linear classifiers are shown at the left side. The right side shows one plausible (green square) and one unlikely (red square) hair hypothesis, with the corresponding pairwise costs.

1556-6013 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TIFS.2017.2680246, IEEE Transactions on Information Forensics and Security

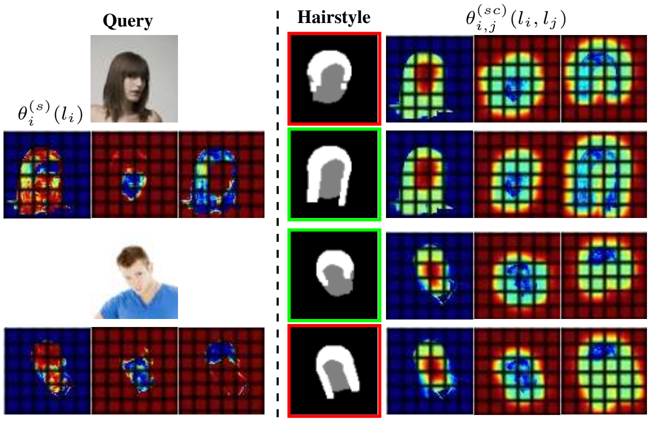

Fig. 5. Rationale for the inter-layer pairwise costs $\theta^{(sc)}_{i,j}(l_i, l_j)$. For two queries, the unary costs (segmentation layer) are shown at the left side, At the right side, having one plausible (green frame) and one non-acceptable (red frame) hair model, the inter-layer pairwise costs encode the reasonability of fitting the data appearance term to the corresponding models (warm colors denote high costs). During inference, the MRF converges into an equilibrium between $\theta^{(sc)}_{i,j}(l_i, l_j)$ and $\theta^{(s)}_{i}(l_i)$.

TABLE II

COMPARISON BETWEEN THE PIXEL SEGMENTATION PERFORMANCE (AFLW SET). THE SUPERSCRIPTS GIVE THE 95% CONFIDENCE INTERVALS.

Method	Overall	Hair	Skin	Background
Proposed Method	$2.87^{\pm 0.03}$	$1.99^{\pm 0.04}$	$3.02^{\pm 0.03}$	$2.95^{\pm 0.03}$
Krupka *et* al. [16]	$4.21^{\pm 0.05}$	$4.05^{\pm 0.06}$	$4.70^{\pm 0.06}$	$4.16^{\pm 0.04}$
Lee *et* al. [17]	$4.23^{\pm 0.05}$	$4.19^{\pm 0.06}$	$5.26^{\pm 0.07}$	$4.09^{\pm 0.04}$
Kae *et* al. [12]	$2.85^{\pm 0.03}$	$2.02^{\pm 0.04}$	$2.98^{\pm 0.04}$	$2.93^{\pm 0.03}$

V. RESULTS AND DISCUSSION

A. Datasets

The LFW [9] was the main dataset used in the empirical validation of our model, due to two reasons: 1) it contains heterogenous images acquired indoor / outdoor, with the degradation factors that are likely in visual surveillance environments; and 2) it has a subset of manually segmented images (the *funnelled* version) into hair, skin and background. Additionally, the AFLW [15] set was considered for evaluating the variations in segmentation performance with respect to errors in the head landmarks detection phase.

B. Model Inference

All our models were optimized using the Loopy Belief Propagation [7] algorithm. Even though it is not guaranteed that it converges to global minimums on non sub-modular graphs (such our models), it provides visually pleasant solutions most of the times. As future work, we plan to evaluate the effectiveness of our method according to more sophisticated energy minimization algorithms (e.g., sequential tree-reweighed message passing [14]).

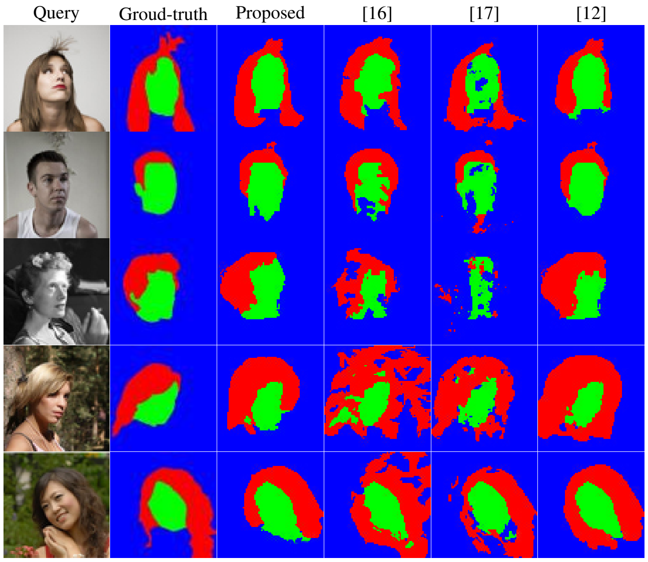

Fig. 6. Typical hair segmentation results obtained by our model (second column), when compared to the methods due to Krupka *et* al. [16] (third column), Lee *et* al. [17] (forth column) and Kae *et* al. [12] (fifth column)

C. Segmentation

We compared the segmentation accuracy of our method to three baseline methods: 1) a computationally inexpensive method due to Krupka *et* al. [29], based on a set of seeds from where the adjacent regions are thresholded; 2) a single layered MRF due to Lee *et* al. [17], which is a particular case of our model, with constant costs in the *objects* layer and in the inter-layer edges; and 3) the method due to Kae *et* al. [12], which we consider the state-of-the-art and has a rationale much similar to our solution: it uses a random field to model the transitions at the pixel level and a restricted Boltzmann machine to enforce globally coherent hypotheses. Fig. 6 illustrates the typical outputs provided by the methods compared: whereas ours and Kae *et* al. methods typically produce similar results, Krupka *et* al. and Lee *et* al. methods are frequently trapped in local minima of their cost functions, due to not enforcing the biological coherence of the solutions. Particularly, Krupka *et* al. produce poor results when the seeds do not faithfully represent the distributions of the components (due to textured data). Finally, by regarding exclusively image appearance, Lee et al.'s method often produces biological unlikely solutions, with discontiguous skin / hair regions with boundaries having too many number of degrees-of-freedom.

More objectively, Table II quantifies the average segmentation performance for the methods evaluated. We got slightly better results than Kae *et* al. for the hair component and worse results for the skin and background, in all cases with differences not being statistically significant (inside the 95% confidence intervals). The method due to Krupka *et* al. ranked third, yet it was the one that most frequently produced biologically inconsistent solutions. Also, this method performed particularly poor in highly textured background images, where the seeds hardly represent the high entropy in the background regions.

Another interesting feature of our method is its ability to

1556-6013 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TIFS.2017.2680246, IEEE Transactions on Information Forensics and Security

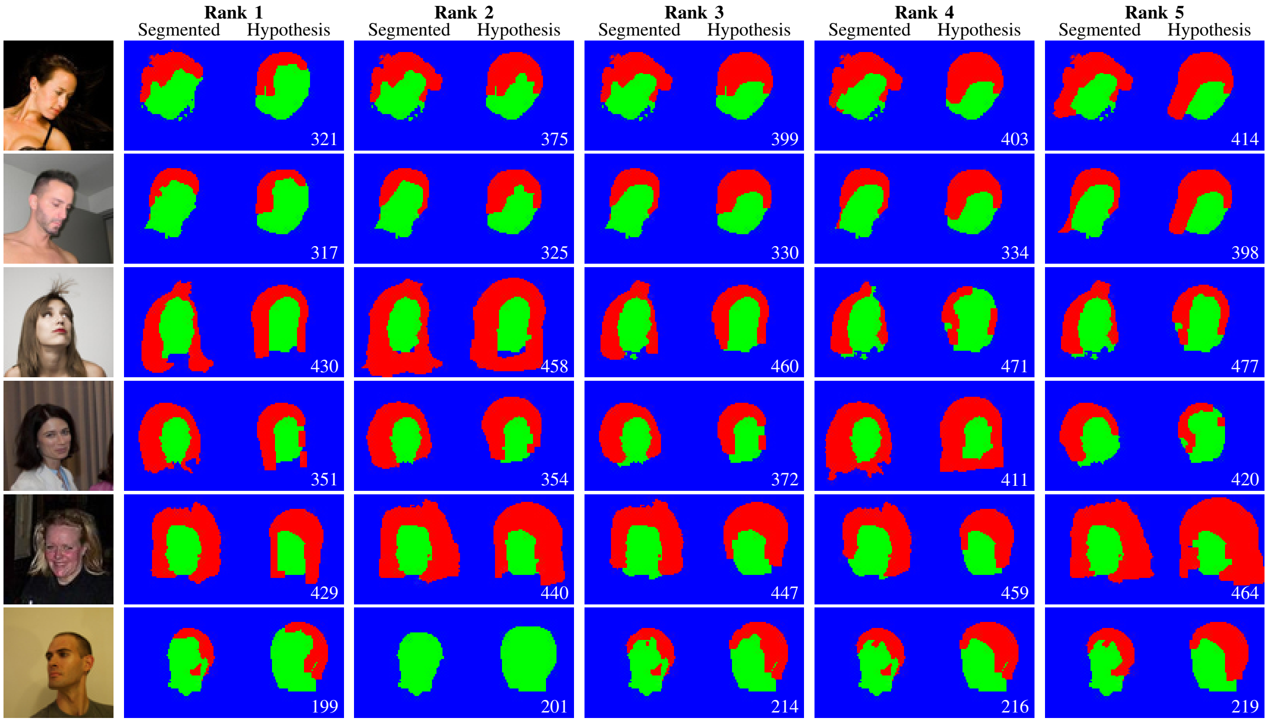

Fig. 7. Hair segmentation / hairstyle inference results obtained by the proposed method. The queries are shown at the leftmost column. For each query, the most likely segmentation and hairstyle models are given in descending likelihood order (from left to right), showing also the cost of the optimal MRF state.

rank the plausibility of the hypotheses with respect to the queries. This can be done by optimizing the model iteratively and, at each step, remove the hypothesis considered optimal in the previous iteration. Results of this ordering are shown in Fig. 7, with the top-5 most similar hair hypotheses with respect to queries, along with the segmentation masks for each hypothesis. At the bottom-right corner, the cost of the solution is given, i.e., the cost of fitting the segmentation mask in the corresponding template.

Note that our results depend of the head landmarks to infer the head shape and pose hypotheses (Sec. III-C). Failures at this point introduce a bias in the way hypotheses are projected and in the MRF unary / pairwise costs. Hence, we used a set of ground-truth head landmarks (AFLW set) to perceive the sensitivity to this factor, introducing inaccuracies in landmarks detection by adding random offsets to the accurate landmarks. Results are given in Fig. 8 (the overall accuracy is shown), with respect to the proportion of landmarks inaccurately detected (horizontal axis) and the relative magnitude of the offset (i.e., the Euclidean distance between the original and the biased landmark positions, weighted by the image diagonal length, vertical axis). The segmentation performance remained approximately invariant when less than 20% of the head landmarks were inaccurate. Also, the magnitude of the detections offset was observed to play a relatively minor role in segmentation accuracy, but, in practice, the algorithm looses its effectiveness when more than 35% of the landmarks are inaccurate.

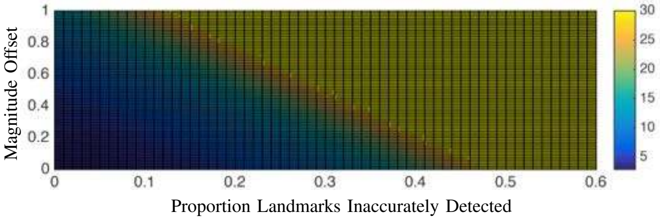

Fig. 8. Variations in segmentation performance of the proposed method with respect to the proportion of head landmarks inaccurately detected (horizontal axis) and the magnitude of these inaccuracies (vertical axis). The overall accuracy is shown.

D. *Identity Retrieval*

This section reports the identity retrieval results in the LFW set. A one-dimensional manifold $\mathbf{M}$ for the hair models was inferred using a self-organized map fed by a feature set composed of the concatenation of the mode label in local 3D volumes regularly sampled in the 3D hair models $\mathbf{S}_h$: $\mathbf{M}$:={0: bald, 1: short bald, 2: short, 3: medium, 4: long fine, 5: long volume}. Let $\mathbf{I}_j^{(i)}$ be the j^{th} image from the i^{th} subject ($j = 1, \dots t_i$), with t_i representing the number of images for that subject. Let $\varepsilon(I_{\cdot}^{(\cdot)})$: $\mathbb{N}^2 \rightarrow \mathbf{M}$ be the inference function (MRF) that associates one query to one hair style in $\mathbf{M}$. $|\varepsilon(I_j^{(i)}) - \varepsilon(I_k^{(i)})|_1$ captures the spread of the intra-subject labels distribution, with the probability density function for this value shown in the upper part of Fig. 9. Results are given with respect to the κ_2 parameter that controls the shape of the transfer function (Sec. IV-B2). In all cases, it is obvious

1556-6013 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TIFS.2017.2680246, IEEE Transactions on Information Forensics and Security

that large deviation values (> 3, for $\kappa_2 = 2.0$) in intra-subject labels rarely occur, which is the insight for using these labels in identity retrieval. The bottom plot gives the corresponding hit / penetration plots, once again with respect to the κ_2 parameter.

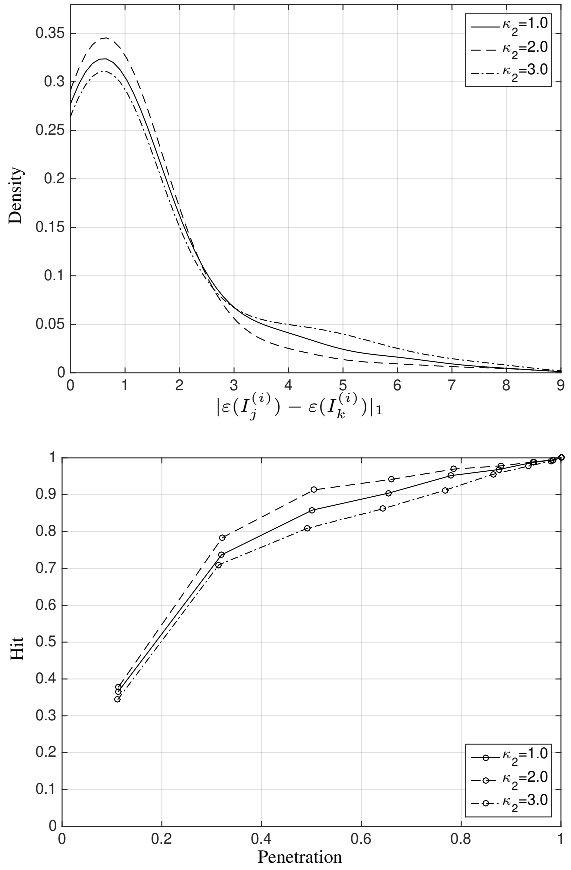

Fig. 9. Top plot: probability density functions for the distance between intra-subject labels $|\varepsilon(I_j^{(i)}) - \varepsilon(I_k^{(i)})|_1$. Bottom plot: hit / penetration plots for the LFW data set.

VI. Conclusions

Being a classical tool in computer vision, MRFs traditionally have difficulties in assuring globally coherent solutions without using too-high order cliques that compromise the computational effectiveness of the inference process. In this paper we described a hierarchical architecture for MRFs free of high-order cliques that still enforces globally coherent models. The idea is to have the bottom layer working at the local (pixel) level, while the upper layers work at the hypotheses level, providing possible solutions for the problem. During optimization, all layers interact and converge into an equilibrium state, where the configuration in the bottom layer implicitly segments the data, and the configuration in the other layers correspond to the most likely models. As test case, we considered the segmentation and labelling of hair / facial hair styles in degraded data, which are important soft biometric labels for human recognition *in-the-wild*. Our experiments were carried out in the challenging LFW data set, and we observed performance similar to the state-of-the-art methods, both in the hair segmentation and hairstyle labelling tasks, and at a much lower computational cost. Further, the proposed MRF architecture can be applied with minimal adaptations to other segmentation / classification computer vision problems, particularly in cases where the biological (global) coherence of the solutions can be objectively measured.

References

[1] A. Albiol, L. Torres, E. Delp. Optimum color spaces for skin detection. In Proceedings of the *International Conference on Image Processing*, vol. 1, pag. 122–124, 2001.

[2] A. Besbes, N. Komodakis, G. Langs, N. Paragios. Shape Priors and Discrete MRFs for Knowledge-based Segmentation. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, pag. 1295–1302, 2009.

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

[4] B. Caputo, S. Bouattour, H. Niemann. Robust appearance-based object recognition using a fully connected Markov random field. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 3, pag. 565–568, 2002.

[5] J. Dass, M. Sharma, E. Hassan, 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*, pag. 1–4, 2013.

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

[7] P. Felzenszwalb, D. Huttenlocher. Efficient Belief Propagation for Early Vision. *International Journal of Computer Vision*, vol. 70, no. 1, pag. 41–54, 2006.

[8] B. Glocker, D. Zikic, N. Komodakis, N. Paragios, N. Navab. Linear Image Registration Through MRF Optimization. In Proceedings of the *IEEE International Symposium on Biomedical Imaging*, pag. 422–425, 2009.

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

[10] S. Jaiswal, T. Almaev, 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.

[11] P. Julian, C. Dehais, F. Lauze, V. Charvillat, A. Bartoli, A. Choukron. Automatic Hair Detection in the Wild. In Proceedings of the *2010 International Conference on Pattern Recognition*, pag. 4617–4620, 2010.

[12] A. Kae, K. Sohn, H. Lee, E. Learned-Miller. Augmenting CRFs with Boltzmann Machine Shape Priors for Image Labeling. In Proceedings of the *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pag. 2019–2026, 2013.

[13] Z. Kato, T.C. Pong. A Markov random field image segmentation model for textured images. *Image and Vision Computing*, vol. 24, pag. 1103–1114, 2006.

[14] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pag. 1568–1583, 2006.

[15] M. Koestinger, P. Wohlhart, P. Roth, 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 Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[16] A. Krupka, J. Prinosil, K. Riha, J. Minar. Hair Segmentation for Color Estimation in Surveillance Systems. In Proceedings of the *Sixth International Conference on Advances in Multimedia*, pag. 102–107, 2014.

[17] K. Lee, D. Anguelov, B. Sumengen, S. Gokturk. Markov Random Field Models for Hair and Face Segmentation. in Proceedings of the *2008 IEEE International Conference on Face and Gesture Recognition*, pag. 1–6, 2013.

[18] U. Lipowezky, O. Mamo, A. Cohen. Using Integrated Color and Texture Features for Automatic Hair Detection. In Proceedings of the *IEEE Convention of Electrical and Electronics Engineers in Israel*, pag. 51–55, 2008.

1556-6013 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TIFS.2017.2680246, IEEE Transactions on Information Forensics and Security

[19] H. Proença, João C. Neves, Silvio Barra, Tiago Marques, Juan C. Moreno. Joint Head Pose / Soft Label Estimation for Human Recognition In-The-Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2016.2522441, 2016. (in press)

[20] H. Proença, João C. Neves, G. Santos. Segmenting the Periocular Region using a Hierarchical Graphical Model Fed by Texture / Shape Information and Geometrical Constraints. In Proceedings of the *International Joint Conference on Biometrics*, pag. 1–7, 2014.

[21] V. Rapp, T. Senechal, K. Bailly, 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.

[22] C. Rousset, P. Coulon. Frequential and color analysis for hair mask segmentation. In Proceedings of the *International Conference on Image Processing*, pag. 2276–2279, 2008.

[23] Y. Shen, Z. Peng, Y. Zhang. Image Based Hair Segmentation Algorithm for the Application of Automatic Facial Caricature Synthesis. *The Scientific World Journal*, vol. 2014, doi: 10.1155/2014/748634, 2014.

[24] P. Tome, J. Fierrez, R. V-Rodriguez, 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.

[25] Y. Ugurlu. Head Posture Detection Using Skin and Hair Information. In Proceedings of the *2010 International Conference on Pattern Recognition*, pag. 1–4, 2012.

[26] C. Wang, M. Gorce, N. Paragios. Segmentation, Ordering and Multi-Object Tracking using Graphical Models. In Proceedings of the *in 12th International Conference on Computer Vision*, pag. 747–754, 2009.

[27] N. Wang, H. Ai. Hair Style Retrieval by Semantic Mapping on Informative Patches. In Proceedings of the *First Asian Conference on Pattern Recognition*, pag. 110–114, 2011.

[28] D. Wang, S. Shan, H. Zhang, W. Zeng, X. Chen. Isomorphic Manifold Inference for Hair Segmentation. In Proceedings of the *2013 IEEE International Conference on Face and Gesture Recognition*, pag. 1–6, 2013.

[29] Y. Wang, Z. Zhou, E. Teoh, B. Su. Human Hair Segmentation and Length Detection for Human Appearance Model. In Proceedings of the *2010 International Conference on Pattern Recognition*, pag. 450–454, 2014.

[30] Y. Yacoob, L. Davis. Detection and Analysis of Hair. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pag. 1164–1169, 2006.

[31] J. Young. Head and Face Anthropometry of Adult U.S. Civilians. Office of Aviation medicine, Federal Aviation Administration, DOT/FAA/AM-93/10, 1993.

[32] C. Yuksel, S. Schaefer, J. Keyser. Hair Meshes. *ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH Asia*, vol. 28, no. 5, article no. 166, 2009.

[33] Z. Zhang, H. Gunes, M. Piccardi. An Accurate Algorithm for Head detection Based on XYZ and HSV hair and Skin Models. In Proceedings of the *International Conference on Image Processing*, pag. 1644–1647, 2008.

[34] Z. Zhang, H. Gunes, M. Piccardi. Head Detection for Video Surveillance Based on categorical Hair and Skin Colour Models. In Proceedings of the *International Conference on Image Processing*, pag. 1137–1140, 2009.

Hugo Proença (SM'12) 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 Boards of the Image and Vision Computing and International Journal of Biometrics journals 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 (M'15) 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.

1556-6013 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.