Biometric Recognition in Surveillance Environments Using Master-Slave Architectures

Hugo Proença
University of Beira Interior
IT: Instituto de Telecomunicações
Covilhã, Portugal 6200–001
hugomcp@di.ubi.pt

João C. Neves
Tomiworld®
Viseu, Portugal 3500–106
JoaoNeves@tomiworld.com

***Abstract*—The number of visual surveillance systems deployed worldwide has been growing astoundingly. As a result, attempts have been made to increase the levels of automated analysis of such systems, towards the reliable recognition of human beings in fully covert conditions. Among other possibilities, master-slave architectures can be used to acquire high resolution data of subjects heads from large distances, with enough resolution to perform face recognition. This paper/tutorial provides a comprehensive overview of the major phases behind the development of a recognition system working in outdoor surveillance scenarios, describing frameworks and methods to: 1) use coupled wide view and Pan-Tilt-Zoom (PTZ) imaging devices in surveillance settings, with a wide-view camera covering the whole scene, while a synchronized PTZ device collects high-resolution data from the head region; 2) use soft biometric information (e.g., body metrology and gait) for pruning the set of potential identities for each query; and 3) faithfully balance ethics/privacy and safety/security issues in this kind of systems.**

I. ”Surveillance” and ”Biometric Recognition” Settings → ”Non-Cooperative Recognition”

Various attacks in crowded urban environments have been reducing the perception of safety in modern societies, while the citizens' tolerance to what they recognize as *reasonable* risks has also been decreasing. There are growing needs of assuring the safety of people, specially in places/events that concentrate large crowds, which are perceived as those with the highest risk (e.g., due to 2001 New York 9/11, 2004 Madrid train bombing or 2013 Boston marathon attacks).

To counterbalance these fears, the use of visual surveillance has been promoted worldwide. The deployment of outdoor surveillance cameras has grown astonishingly in the recent years, with more than 5.9 million CCTV cameras already set in the United Kingdom [3]. However, the automated understanding of data in this kind of systems is still mostly reduced to action recognition. Simultaneously, biometrics is considered an especially successful case of pattern recognition: systems have been deployed for different applications (e.g., security assess or refugee control), but performance is still strongly conditioned by the levels of cooperation demanded from subjects and by the environmental conditions required to obtain data with minimal levels of quality.

There is an evident complementarity between the ’visual surveillance” and ”biometrics” domains, not only in the environmental conditions under which both kinds of systems work, but also the tasks they perform: surveillance systems work in uncontrolled conditions but do not automatically identify suspects in a crowd, whereas biometric systems are effective in automatic identification, but work in environments that enable to acquire good quality data [38].

According to the above points, the tutorial described in this paper focuses on the research carried out to develop biometric recognition systems that work in conditions currently associated to visual surveillance. In particular, we discuss the challenges behind master-slave architectures, composed of: 1) a wide-view (static) camera covering the whole outdoor scene; and 2) a Pan-Tilt-Zoom (PTZ) device that points to specific regions in the scene and acquires data of the human head, with enough resolution to be used in biometric recognition (Fig. 1).

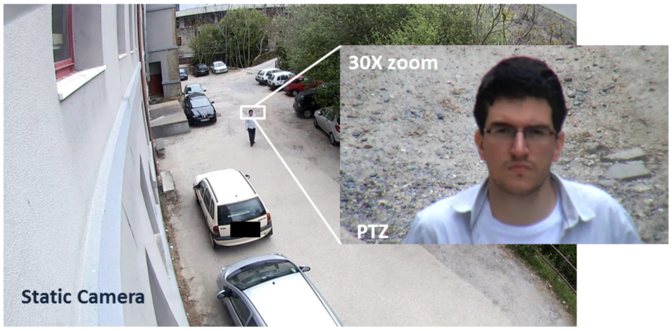

Fig. 1. Example of the data resulting from a master/slave dual camera architecture for biometric recognition in surveillance environments: while a wide view (static) camera covers the whole scene, the coordinates of detected human heads are sent to a calibrated Pan-Tilt-Zoom device that acquires data with enough resolution to be used in biometric recognition.

II. Master-Slave Data Acquisition Architectures

A. System Overview

In terms of the hardware infrastructure required by master/slave visual surveillance recognition systems, Fig. 2 gives an example of a laboratorial prototype mounted at the outdoor wall of our lab[1]. The rationale is to use the PTZ camera as a foveal sensor, i.e., the video stream from the wide camera is analyzed to infer the location of subjects' heads, so that

[1] SOCIA: Soft Computing and Image Analysis Lab., http://socia-lab.di.ubi.pt

the PTZ camera can acquire samples of the facial regions at a high-magnification state.

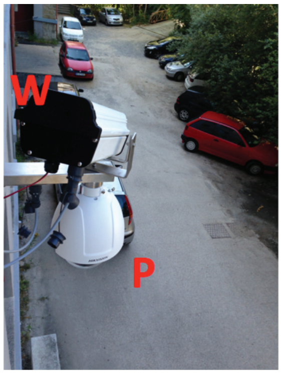

Fig. 2. Example of a synchronised pair of wide-view (W) and PTZ (P) cameras, from a master/slave laboratorial prototype mounted at the *SOCIA: Soft Computing and Image Analysis* laboratory outdoor wall. This prototype is able to automatically detect and track subjects passing by and acquire high resolution images of subjects heads located up to 50 meters away of the recognition system.

Regarding the software components, they can be broadly divided into: 1) low-level vision tasks; and 2) high-level vision tasks. The first group contains the phases that are required to run in real-time and involve the perception of the scene, up to the moment the PTZ device is pointed out to a particular 3D position, in order to acquire a high-resolution image of a subject's head. All the subsequent phases belong to the second group and don't have strict requirements in terms of the computational burden, as they can be forked to separate processes, each one performing for one recognition attempt per query sample.

B. Camera Calibration

The requirement of inter-camera calibration is the major bottleneck of master/slave configurations, since determining the mapping function from static image coordinates to pan-tilt parameters requires depth information. To address this problem, most master-slave systems use 2D-based approximations, but, in turn, they are compelled to rely on different assumptions (e.g., similar points-of- view or intermediate zoom states) to alleviate pan-tilt inaccuracies. The use of multiple optical devices has been pointed as a solution to infer depth information through triangulation. Choi *et* al. [8] and Park *et* al. [43] were the first to exploit this alternative without using stereographic reconstruction, which is not feasible in real-time applications. Instead, they disposed the cameras in a coaxial configuration to ease triangulation. In addition, the authors ascertained the feasibility of facial recognition at-a-distance using the proposed calibration method. However, the highly stringent disposition of the cameras restrains its use in outdoor environments as well as its operational range (up to 15m)

Aiming at improving the existing master-slave systems, in particular the work of Choi *et* al. [8] and Park *et* al. [43], we extended PTZ-assisted facial recognition to surveillance scenarios. We introduced [36] a calibration algorithm capable of accurately estimating pan-tilt parameters without resorting to intermediate zoom states, multiple optical devices or highly stringent configurations. Our approach exploits geometric cues (the vanishing points available in the scene) to automatically estimate subjects height and thus determine their 3D positions (Fig. 3). Furthermore, we have built on the work of Lv *et* al. [30] to ensure robustness against human shape variability during walking.

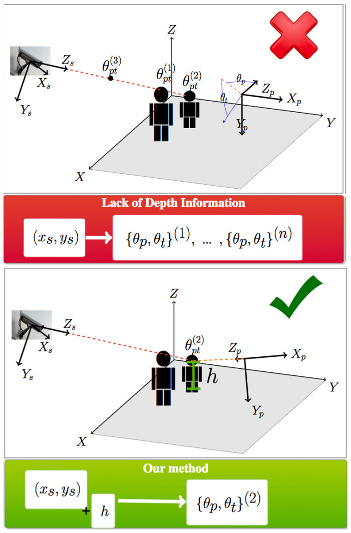

Fig. 3. Illustration (adapted from [36]) of the principal bottleneck of master-slave systems and the proposed strategy to address this problem. One image pixel (x_s, y_s) might correspond to multiple 3D positions in the scene, and consequently to different pan-tilt $\{\theta_p, \theta_t\}$ values. Our work is based on the premise that human height h can be exploited to infer depth information and avoid the 2D → 3D ambiguity.

III. Low-Level Vision Tasks

This section describes the major phases that compose the initial part of the processing chain: from the global analysis of the scene up to the moment when the PTZ device is pointed to image a subject's head. We start by pruning the data coming from the wide-view camera (background subtraction), followed by human detection and tracking steps. Next, the order for imaging the existing subjects is established and their corresponding coordinates are sent to the PTZ device.

A. Background Subtraction

Background subtraction (BS) aims at dividing the scene into two disjoint parts: 1) the background, containing the static regions in the input data that should be disregarded; and 2) the foreground, that contains the regions-of-interest (ROIs) of the objects that the system should care about. Essentially, this phase enables to prune the scene and reduce the amount of information to be handled [40].

The existing methods for background subtraction can be divided into three families: 1) basic; 2) Gaussian-based; and 3) machine-learning based, ordered by their level of complexity. The first family regards the most simple strategies and the pioneering approach dates back to 1979, when Jain and Nagel [18] analysed the differences in pixel intensity with respect to time to discriminate the useless regions in a scene. Similar approaches were published (e.g., [60]) and have as main advantage their reduced computational cost. Other type of methods derives a coarse estimate of the background from the earliest frames or from the last frames in a scene, using simple statistics with respect to time (median filtering).

A family of methods of intermediate complexity models the density of the intensities of each pixel with respect to changes in time, either assuming single Gaussian (e.g., [67]), mixture of Gaussians (e.g., [61]) or non-parametric models (e.g., [13]). The underlying strategy is to obtain an estimate of the typical changes in each position in the image, and then report a pixel as foreground every time an outlier is observed. Simultaneously, the models are continuously updated, so to adapt to slight changes in the scene.

The most complex family of methods relies in machine-learning algorithms to obtain local representations of the background. Clustering-based approaches estimate the background by grouping pixels in different clusters, each one corresponding to a different source of background. The Codebook model [22] uses a set of words to represent each cluster. Nearest neighbour techniques are used for background estimation, fed by features such as luminance and chrominance [6]. More recently, unsupervised neural models have been tested to enhance robustness in real-world conditions. In this kind of methods, each pixel is modelled by a neural map, where each element stores typical RGB values at that position and acts as cluster centroid,. The idea behind competitive neural networks [29] is highly similar to this strategy.

B. Human Detection

The detection of humans can be seen simply as a particular instance of object detection, with some specificities that increase the difficulty of the task. Such challenges include the deformations in shape of the human silhouette with respect to movements of the legs, arms and head, partial occlusions, clothing and changes in perspective. Using the output of the background subtraction module, the goal is to define a set of regions-of-interest (ROIs), such that each one corresponds to one of the humans in the scene. From the computer vision perspective, working in outdoor environments, with both global and local variations in lighting conditions, and imposing no constraints about the number of subjects represents a highly challenging problem, that can be even harder in case of poor data resolution [69].

There are two major families of methods for human detection: 1) holistic methods, where the whole body is searched at once in the image; and 2) part-based methods, where each part of the body is detected independently, with the information being further fused for consistency purposes. Most of the holistic methods learn a discriminative model, being well represented by the popular Viola and Jones' [64] method, adapted to detect humans using motion patterns [65]. Similarly, Dalal and Triggs [10] use histograms of oriented gradients (HOGs) to feed a classifier, such as support vector machines (SVMs), in a way similar to [33]. Along with HOGs, local binary patterns have also been widely used for human detection (e.g., [66]).

Regarding part-based methods, Mikolajczyk *et* al. [32] use a probabilistic model to assemble all parts of the body, each one detected in a corse-to-fine strategy. Lin *et* al. [26] considered the head the most reliable part to be detected and to estimate the number of persons in a large crowd, similarly to Subburaman *et* al. [62]. Zhao and Nevatia [71] analysed the silhouette boundaries from the background estimation mask and detect the head by searching for vertical peaks on these contours. Wu and Nevatia [68] use four body parts (full-body, head-shoulder, torso, and legs), each one learned by boosting a set of weak classifiers based on edgelet features (short segments of edge pixels). The responses given by all detectors are fused to provide robustness to occlusions.

C. Tracking

Given an initial estimate of the location of one object, the tracking phase aims at determining the positions of that object in the subsequent frames, having typically two major goals: 1) by perceiving the object path, accurate predictions of the location of the object in forthcoming frames can be made, which allows to timely point the PTZ device for a specific position; and 2) once the high-resolution data of a subject is acquired, that object can be ignored of any subsequent processing.

Approaches for object tracking can be divided with respect to the feature space they work in: 1) motion-based algorithms, which exploit the object dynamics based on cues such as velocity, articulation and periodic constraints. Motion models are typically related to Bayesian tracking approaches, where dynamics is used to update the target state over time (e.g., Breitenstein *et* al. [5]) or shape information (e.g., Zhou *et* al. [74]). Tracking based on optical flow estimation is also a relevant example of this family, namely the KLT tracker [59], that assumes small movements between frames with brightness constancy, to follow a set of keypoints; 2) appearance-based algorithms are frequently associated to kernel-based methods that represent the target as a point in a high dimensional space, characterized in terms of histograms of intensities (e.g., Comaniciu *et* al. [9]), LBPs (e.g., Kalal *et* al. [21]) or sparse representations (e.g., Zhong *et* al. [72]) from channels of different color spaces; 3) shape-based algorithms eliminate the need to consider varying illumination and changes in appearance, yet turn more difficult to obtain a reliable estimate of the object boundary. However, shape information is most times used together with other families of cues (e.g., texture), which is particularly useful for low quality data (e.g., Liu *et* al. [27]).

Complementary, tracking methods can be classified with respect to the properties of their main algorithm. The earliest approaches attempted to track objects by searching for specific patterns in the neighbourhood of the previous known location (kernel / model tracking) or by evolving the state of the target according to a motion and appearance model (Bayesian tracking). More recently, a new strategy has been gaining popularity (tracking-by-detection), which is particularly suitable for arbitrary object tracking in unconstrained scenarios. A typical example of this family of algorithms is the proposal of Zhou and Aggarwal [73], using the Kalman filter with a constant velocity model to estimate the state of humans. Aiming at improving the robustness of tracking to dynamic environments, Zhang *et* al. [70] use a kernel-based Bayesian framework, where the feature space combines appearance and shape information. Mixture of Gaussians are also used to model the appearance model and the Chamfer matching provides a similarity measure between shapes. In case of shape cues, it is particular hard to be match shapes subject to severe occlusions and deformed shapes. For this reason, Saber *et* al. [58] use the concept of partial shape matching, as Husain *et* al. [16] did, to track objects in surveillance scenarios.

As stated above, the iterative use of detectors has been gaining popularity, mainly due to the high flexibility of this kind of algorithms and the hardware advances that have been reducing the amount of time required for execution. These algorithms estimate the target position by searching the position in the image that maximizes a similarity function between an image point and the feature vector of the target state. Contrary to other tracking families, no a priori target representation is required, inferring the corresponding model by online learning algorithms, allowing the resulting model to adapt to any kind of object and its variations in appearance. Regarding the used machine-learning classifier, online boosting classifiers were a typical strategy in the earliest approaches, but state-of-the-art techniques exploit multiple instance learning techniques to reduce the sensitiveness to slight changes in appearance (e.g., Babenko *et* al. [2]).

Recently, substantial attention has been paid to multiple object tracking. Despite multiple instances of a tracking algorithm can be used to address multiple targets, there is an exponential growth of computational complexity that restrains their use when the number of targets is high. Greedy strategies have been used to handle such complexity, where correspondences are regarded as an assignment problem based on spatial distance (e.g., Wu and Nevatia [68]). Offline or batch techniques methods comprise another solution for multiple target tracking, using the complete set of detections before estimate the trajectory. This phase is regarded as an optimization problem, where a function describes the cost of each solution. Linear programming techniques are used in several works to cope with the computational burden of this optimization step. A continuous formulation of this problem was introduced by Andriyenko and Schindler [1], which has as main drawback the high latency required to analyse a video, that turns it incompatible with real-time requirements.

D. Camera Scheduling: Target Selection

Camera scheduling is the process that determines the order by which the targets (subjects) in a scene will be imaged, given a set of scene features (e.g., subject pose, distance, velocity, levels of occlusion and number of previous imaging attempts per subject). Scheduling in PTZ-based systems can be broadly divided in coverage and saccade approaches. In the former, the cameras are set in an intermediate zoom state so that multiple targets are observable by the same device. The goal is to maximise the number of targets seen by the complete set of cameras. On the contrary, in a saccade approach each camera just observes one target at a time. A sequence of saccades is planned, in real-time, to maximize the number of different targets observed and minimize the cumulative transition time. Even though previous works have presented solutions to variants of this problem, but Costello *et* al. [7] were the first to explicitly define and propose a solution to this problem. Considering the similarities with network packet routing problem, the authors proposed the use of the Current Min loss Throughput Optimal method to schedule a set of observations. A similar strategy was used by Qureshi and Terzopoulos [54], where a greedy best-first search was employed to determine the optimal plan. Previously, Qureshi and Terzopoulos [53] have relied on greedy algorithms such as the Shortest Elapsed Time First and weighted Round Robin (RR) for the same purpose. Krahnstoever *et* al. [24] discussed the best heuristics to dynamically estimate new observation plans. Targets were modelled as graph nodes and transition costs were defined according to their distance to the camera and expected time to exit the scene. Lim *et* al. [25] constructed a directed acyclic graph based on the starting time of ”task visibility intervals”, which were inferred by prediction. The scheduling problem was formulated as a maximal flow problem and a dynamic programming scheme was proposed to solve it. Ilie and Welch [17] relied on a greedy algorithm to determine a plan based on geometric reasoning.

Based on the above works, we came out with a proposal [37] for dynamic camera scheduling, capable of determining - in real-time - the sequence of acquisitions that maximizes the number of different targets obtained, while minimizing the cumulative transition time. Our approach models the problem as an undirected graphical model (Markov Random Field, MRF), as illustrated in Fig. 4, which energy minimization can approximate the shortest tour to visit the maximum number of targets.

IV. High-Level Vision Tasks

The low-level group of tasks is responsible for acquiring high resolution images of moving subjects passing by in the scene up to 50 meters away from the acquisition system. This is the kind of data that feeds the high-level vision processing phases, which are responsible for performing biometric recognition. Fig 5 gives some examples of human head samples acquired by our prototype. In our case, the recognition step was further divided into two parts: 1) aiming at pruning the set of plausible identities for a given query, which was done mostly according to soft biometric information; and 2) aiming at establish an unique correspondence between a sample and a known identity, according to the face (e.g. [39]), the iris (e.g., [44], [47] and [52]), the periocular

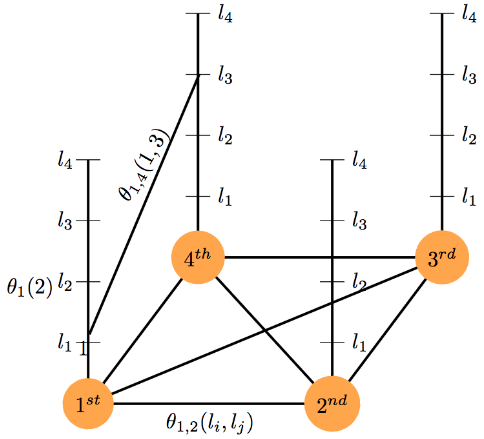

Fig. 4. Illustrative example (taken from [37]) of the MRF used when four targets are in the scene. Labels encode the set of targets in the scene, whereas the nodes correspond to the order that they will be imaged. Unary costs represent the angular differences between the current position of the camera and each target, whereas pairwise costs model the transition angles.

(e.g., [46], [45], [45], [48] and [34]), or the gait (e.g. [42]) traits.

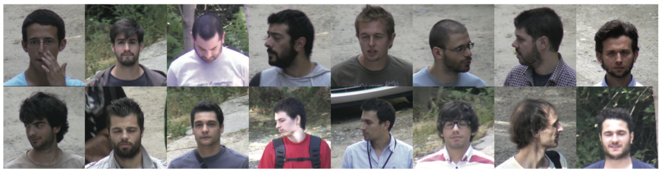

Fig. 5. Examples of images (taken from [49]) acquired by a visual surveillance system, composed by a wide-view camera feeding a pan-tilt-zoom device that collects data from moving and at-a-distance targets (up to 50 meters away).

A. Identities Pruning: Soft Biometrics

According to [63], 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 subjects 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. [14] 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 [19] used demographic information (gender and ethnicity) and facial marks (scars, moles and freckles) to improve face image matching and retrieval performance.

In terms of body traits, Lucas and Henneberg [28] 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. [57]) concluded that identification based on body measurements can be as accurate as using the face. Moustakas *et* al. [35] 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. [12] 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, various works analyze the discriminability of hair/facial hair styles and lengths. Dass *et* al. [11] pre-aligned the images based on the position of the eyes and defined five groups of hairstyles according to hair density in image patches. Hewig *et* al. [15] 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. [56]: *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. [23] and Reid *et* al. [55].

B. Identities Pruning: Head Shape

Analysing high resolution images of the subjects heads (illustrated in Fig. 5) enables to infer not only the head poses but also to obtain a coarse estimation of the head shape of the subjects, which can be used as auxiliary soft biometric information. According to this idea, we proposed a method to infer jointly human head poses and soft labels [49].

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 , 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 soft labels yield from the head shape hypothesis that is assigned to each sample. Based on two dimensional manifolds of head shapes (illustrated in Fig. 6), that intrinsically represent *head shape similarity*, we obtain a topologically ordered space, i.e., neighbor prototypes feature similar head shapes. Later, 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.

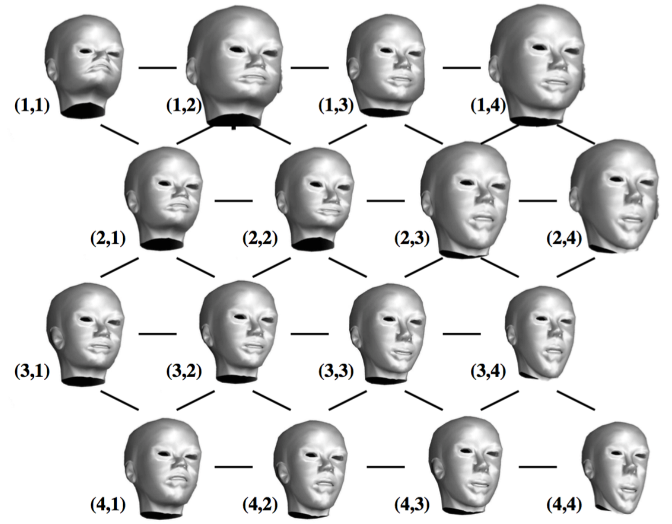

Fig. 6. Representation of the 3D head centroids resulting of a 4 × 4 manifold (taken from [49]). Note the similarity in size / shape between adjacent elements, rooted in the preservation of the topological properties of the input space that this kind of maps offers.

C. *Identities Pruning: Hair Analysis*

The descriptions of the facial hair and hair styles are among the most effective soft biometric traits reported in the literature [63]. 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, the ambition of working effectively in images acquired in typical visual surveillance conditions remains to be achieved.

Having this problem in mind, we proposed a multi-layered (hierarchical) MRF that does not use high order cliques, but still typically reaches globally coherent solutions [50]. We 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; and 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., [51]).

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. Among other advantages, 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.

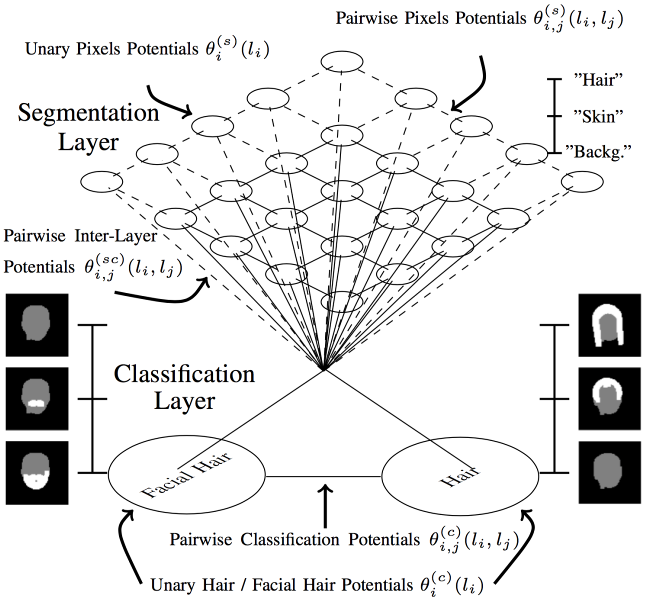
Pairwise Pixels Potentials $\theta^{(s)}_{i,j}(l_i, l_j)$
Unary Pixels Potentials $\theta^{(s)}_{i}(l_i)$
Segmentation Layer
"Hair"
"Skin"
"Backg."
Pairwise Inter-Layer Potentials $\theta^{(sc)}_{i,j}(l_i, l_j)$
Classification Layer
Facial Hair
Hair
Pairwise Classification Potentials $\theta^{(c)}_{i,j}(l_i,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

Fig. 7. Structure of a MRF (taken from [50]) 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).

V. Ethical/Privacy Issues

Undoubtedly, the type of recognition systems discussed here raises serious concerns in terms of the citizens' privacy and of the morality behind recognising someone without asking his permission. There are various national and international laws that regulate the functioning of biometric recognition systems: the *Universal Declaration of Human Rights* contains an article (the 12^{th}), stating that: *"No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks"*. However, the 3^{rd} article of the same declaration also states that *"Everyone has the right to life, liberty and security of person"*. Both articles assure the right to privacy and security, but what happens in cases where rights cannot be assured jointly?

The European Parliament issued a directive (95/46/EC), where the "*notice*" and "*consent*" properties bias the overall policy to forbid non-cooperative recognition systems, while the U.S. government issued a regulation for using biometric data (8 CFR 103.16). These regulations contain subjective statements that make hard to objectively perceive the cases where non-cooperative biometric recognition is acceptable and where it is not. As Benjamim Franklin stated: "*They that can give up essential liberty to obtain a little temporary safety deserve neither liberty nor safety*". On the other side, one should account that the context in B. Franklin's life was completely different, as stated by Neil Young (musician): "*Benjamin Franklin said that anyone who gives up essential liberties to preserve freedom is a fool, but maybe he didn't conceive of nuclear war and dirty bomb*" [4].

At the end, it is to the official entities in each country to consider the levels of threats in each place, and to establish precise limits for the use of non-cooperative surveillance recognition systems.

Acknowledgment

This work is supported by "FCT - Fundação para a Ciência e Tecnologia" (Portugal) through the project UID/EEA/50008/2013. Also, we acknowledge the financial support given by "FEDER, Fundo de Coesão e Fundo Social Europeu", in the scope of the PT2020 - Portugal 2020 program and POCI-01-0247-FEDER-033395 research project.

References

[1] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR'11), pag. 1265–1272, 2011.

[2] B. Babenko, M-H. Yang and S. Belongie. Robust Object Tracking with Online Multiple Instance Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pag. 1619–1632, 2011.

[3] David Barret (The Telegraph). One surveillance camera for every 11 people in Britain, says CCTV survey. http://www.telegraph.co.uk/technology/10172298/, July 10, 2006. (retrieved June 17, 2014)

[4] K.W. Bowyer. Face Recognition Technology and the Security Versus Privacy Tradeoff. *IEEE technology and Society*, pag. 9–20, 2004.

[5] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier and L. Van Gool. Online Multiperson Tracking-by-Detection from a Single, Uncalibrated Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pag. 1820–1833, 2011.

[6] D. Butler, V. Bove and S. Sridharan. Real-Time Adaptive Foreground/Background Segmentation. *EURASIP Journal on Advances in Signal Processing*, vol. 14, 841926, doi: 10.1155/ASP.2005.2292, 2005.

[7] C. J. Costello, C. P. Diehl, A. Banerjee, and H. Fisher. Scheduling an active camera to observe people. In Proceedings of the *ACM 2nd International Workshop on Video Surveillance and Sensor Networks* (WVSSN'04), pag. 39–45, 2004.

[8] H.-C. Choi, U. Park, and A. Jain. Ptz camera assisted face acquisition, tracking and recognition. In Proceedings of the *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems* (BTAS'10), pag. 1–6, 2010.

[9] D. Comaniciu, V. Ramesh and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pag. 564–577, 2003.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Proceedings of the *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'05), vol. 1, pag. 886–893, 2005.

[11] 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.

[12] 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.

[13] A. Elgammal, D. Harwood and L. Davis. Non-parametric Model for Background Subtraction. In Proceedings of the *European Conference on Computer Vision* (ECCV'00), LNCS, vol. 1843, pag. 751–767, 2000.

[14] 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, 1997.

[15] 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.

[16] M. Husain, E. Saber, V. Misic and S.P. Joralemon. Dynamic Object Tracking by Partial Shape Matching for Video Surveillance Applications. In Proceedings of the *International Conference on Image Processing*, (ICIP'06), pag. 2405–2408, 2006.

[17] A. Ilie and G. Welch. Online control of active camera networks for computer vision tasks. *ACM Transactions on Sensor Networks*, vol. 10, no. 2, pag. 25–40, 2014.

[18] R. Jain and H-H. Nagel. On the Analysis of Accumulative Difference Pictures from Image Sequences of Real World Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pag. 206–214, 1979.

[19] 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.

[20] J. Neves, H. Proença Dynamic Camera Scheduling for Visual Surveillance in Crowded Scenes using Markov Random Fields. In Proceedings of the*12th IEEE International Conference on Advanced Video and Signal based Surveillance* (AVSS'15), pag. 1–6, 2015.

[21] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-Learning-Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pag. 1409–1422, 2012.

[22] K. Kim, T. Chalidabhongse, D. Harwood and L. Davis. Real-time foreground background segmentation using codebook model. *Real-Time Imaging*, vol. 11, no. 3, pag. 172–185, 2005.

[23] 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.

[24] N. Krahnstoever, T. Yu, S.-N. Lim, K. Patwardhan, and P. Tu. Collaborative Real-Time Control of Active Cameras in Large Scale Surveillance Systems. In Proceedings of the*Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications* (M2SF2'15), 2008.

[25] S.-N. Lim, L. Davis, and A. Mittal. Task scheduling in large camera networks. In Proceedings of the*of the Asian Conference on Computer Vision* (ACCV'07), pag. 397–407, 2007.

[26] S-F. Lin, J-Y. Chen and H-X. Chao. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 31, no. 6, pag. 645–654, 2001.

[27] Z. Liu, H. Shen, G. Feng and D. Hu. Tracking objects using shape context matching. *INeurocomputing*, vol. 83, pag. 47–55, 2012.

[28] 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.

[29] R. Luque, E. Domnguez, E. Palomo and J. Muoz. A Neural Network Approach for Video Object Segmentation in Traffic Surveillance. In Proceedings of the *International Conference on Image Analysis and Recognition* (ICIAR'08), LNCS, vol. 5112, pag. 151–158, 2008.

[30] F. Lv, T. Zhao, and R. Nevatia. Self-calibration of a camera from video of a walking human. In Proceedings of the *16th International Conference on Pattern Recognition* (ICPR'02), vol. 1, pag. 562–567, 2002.

[31] L. Marchesotti, S. Piva, A. Turolla, D. Minetti, and C. S. Regazzoni. Cooperative multisensor system for real-time face detection and tracking in uncontrolled conditions. *Proceedings of the SPIE*, vol. 5689, 2005.

[32] K. Mikolajczyk, C. Schmid and A.Zisserman. Human Detection Based on a Probabilistic Assembly of Robust Part Detectors. In Proceedings of

the *European Conference on Computer Vision (ECCV 2004)*, LNCS, vol. 3021, pag. 69–82, 2004.

[33] D. Moctezuma, C. Conde, IM. de Diego and E. Cabello. Person detection in surveillance environment with HoGG: Gabor filters and Histogram of Oriented Gradients. In Proceedings of the *IEEE International Conference on Computer Vision Workshops* (CVPRW'11), no. 11, pag. 1793–1800, 2011.

[34] J.C. Moreno, V. B. Surya Prasath, G. Santos, H. Proença. Robust periocular recognition by fusing sparse representations of color and geometry information. *Springer Journal of Signal Processing Systems*, 2015.

[35] 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.

[36] J.C. Neves, J.C. Moreno, H. Proença. Acquiring High-resolution Face Images in Outdoor Environments: A master-slave Calibration Algorithm. In Proceedings of the *IEEE Seventh International Conference on Biometrics: Theory, Applications and Systems* (BTAS'15), pag. 1–6, 2015.

[37] J.C. Neves, H. Proença. Dynamic Camera Scheduling for Visual Surveillance in Crowded Scenes using Markov Random Fields. In Proceedings of the *12th IEEE International Conference on Advanced Video and Signal based Surveillance* (AVSS'15), pag. 1–6, 2015.

[38] J.C. Neves, H. Proença. Biometric Recognition in Surveillance Scenarios: A Survey. *Springer Artificial Intelligence Review*, vol. 46, no. 4, pag. 1–27, 2016.

[39] J.C. Neves, H. Proença. "A Leopard Cannot Change Its Spots": Improving Face Recognition Using 3D-based Caricatures. *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 1, pag. 151–161, 2018.

[40] J.C. Neves, K. Wysoczanska, H. Proença. Evaluation of Background Subtraction Algorithms for Human Visual Surveillance. *International Conference on Signal and Image Processing Applications* (ICSIPA'15), pag. 1–6, 2015.

[41] C. Padole, H. Proença. Compensating for Pose and Illumination in Unconstrained Periocular Biometrics. *International Journal of Biometrics*, pag. 336-359, 2013.

[42] C. Padole, H. Proença. Aperiodic Feature Representation for Gait Recognition in Cross-view Scenarios for Unconstrained Biometrics. *Springer Pattern Analysis and Applications*, 2015.

[43] U. Park, H.-C. Choi, A. Jain, and S.-W. Lee. Face tracking and recognition at a distance: A coaxial and concentric PTZ camera system. *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 10, pag. 1665–1677, 2013.

[44] H. Proença. IRINA: Iris Recognition (even) in Inacurately Segmented Data. In Proceedings of the *Conference on Computer Vision and Pattern Recognition*, pag. 6747–6756, 2017.

[45] H. Proença. Ocular Biometrics by Score-Level Fusion of Disparate Experts. *IEEE Transactions on Image Processing*, vol. 23, no. 12, pag. 5082–5093, 2014.

[46] H. Proença. Deep-PRWIS: Periocular Recognition Without the Iris and Sclera Using Deep Learning Frameworks. *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 4, pag. 888–896, 2018.

[47] H. Proença, L.A. Alexandre. Toward Non-Cooperative Iris Recognition: A Classification Approach Using Multiple Signatures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 4, pag. 607–612, 2007.

[48] H. Proença, J.C. Moreno. Periocular Biometrics: Constraining the EGM Algorithm to Biologically Plausible Distortions. *IET Biometrics*, 2013.

[49] H. Proença, J.C. Neves, S. Barra, T. Marques, J. C. Moreno. Joint Head Pose / Soft Label Estimation for Human Recognition In-The-Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 12, pag. 2444–2456, 2016.

[50] H. Proença, J.C. Neves. Soft Biometrics: Globally Coherent Solutions for Hair Segmentation and Style Recognition based on Hierarchical MRFs. *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pag. 1637–1645, 2017.

[51] H. Proença, J.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* (IJCB'14), pag. 1–7, 2014.

[52] H. Proença, G. Santos. Fusing Color and Shape Descriptors in the Recognition of Degraded Iris Images Acquired at Visible Wavelength. *Elsevier Computer Vision and Image Understanding* vol. 116, pag. 167–178, 2012.

[53] F. Qureshi and D. Terzopoulos. Planning ahead for ptz camera assignment and handoff. *Multimedia Systems* vol 12, no. 3, pag. 269–283, 2006.

[54] F. Qureshi and D. Terzopoulos. Surveillance camera scheduling: a virtual vision approach. In Proceedings of the *International Conference on Distributed Smart Cameras* (ICDSC'09), pag. 1–8, 2009.

[55] 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.

[56] 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.

[57] 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.

[58] E. Saber, Y. Xu and A. Tekalp. Partial shape recognition by submatrix matching for partial matching guided image labelling. *Pattern Recognition*, vol. 6312, no. 10, pag. 1560–1573, 2005.

[59] J. Shi and C. Tomasi. Good features to track. in Proceedings of the *12^{th} IEEE International Conference on Computer Vision and Pattern Recognition* (CVPR'94), pag. 593–600, 1994.

[60] W. Shuigen, C. Zhen and D. Hua. Motion Detection Based on Temporal Difference Method and Optical Flow field. In Proceedings of the *Second International Symposium on Electronic Commerce and Security* (ISECS'09), vol. 2, pag. 8–88, 2009.

[61] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In Proceedings of the *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'99), vol. 2. pag. 246–252, 1999.

[62] B. Subburaman, A Descamps and C. Carincotte. Counting People in the Crowd Using a Generic Head Detector. in Proceedings of the *IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance* (AVSS'12), pag. 470–475, 2012.

[63] 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.

[64] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. in Proceedings of the *ComputerVision and Pattern Recognition Conference* (CVPR'01), vol. 1, pag. I-511–I-518, 2001.

[65] P. Viola, J.C. Platt and C. Zhang. Multiple Instance Boosting for Object Detection. in *Advances in Neural Information Processing Systems*, vol. 18, pag. 1417–1426, 2005.

[66] X. Wang, T.X. Han and S.Yan. An HOG-LBP human detector with partial occlusion handling. In Proceedings of the *IEEE International Conference on Computer Vision* (ICCV'09), pag. 32–39, 2009.

[67] C.R. Wren, A. Azarbayejani, T. Darrell and A.P. Pentland. Pfinder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pag. 780–785, 1997.

[68] B. Wu and R.Nevatia. Detection and Segmentation of Multiple, Partially Occluded Objects by Grouping, Merging, Assigning Part Detection Responses. *International Journal of Computer Vision*, vol. 82, no. 2, pag. 185–204, 2009.

[69] J. Yao and J-M. Odobez. Fast human detection from joint appearance and foreground feature subset covariances. *Computer Vision and Image Understanding*, vol. 115, no. 10, pag. 1414–1426, 2011.

[70] X. Zhang, W. Hu, H. Bao and S. Maybank. Robust Head Tracking Based on Multiple Cues Fusion in the Kernel-Bayesian Framework. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pag. 1197–1208, 2013.

[71] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pag. 1208–1221, 2004.

[72] W. Zhong, H. Lu and M-H. Yang. Robust object tracking via sparsity-based collaborative model. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR'12), pag. 1838–1845, 2012.

[73] Q. Zhou and J.K. Aggarwal. Probabilistic recognition of human faces from video. *Image and Vision Computing*, vol. 24, no. 11, pag. 1244–1255, 2006.

[74] S. Zhou, V. Krueger and R. Chellappa. Object tracking in an outdoor environment using fusion of features and cameras. *Computer Vision and Image Understanding*, vol. 91, no. 12, pag. 214–245, 2003.