Synthetic and privacy-preserving visualization of video sensor network outputs

Carmelo Velardo[1], Claudia Araimo[2], Jean-Luc Dugelay[1]

[1] Eurecom
Multimedia department
Sophia Antipolis, France
{velardo,dugelay}@eurecom.fr

[2] Università "La Sapienza"
Ingegneria dell'informazione, elettronica, e
telecomunicazioni
Roma, Italy
araimo.1224611@studenti.uniroma1.it

***Abstract*—We propose a complete framework for people tracking in video surveillance networks that preserves privacy by mapping people recorded by a video sensor network in the map of the corresponding surveilled areas. Thanks to this approach, it is possible (1) to synthesize multiple video outputs into a unique picture, and (2) to control privacy by offering the possibility to filter information to viewers. A set of physical attributes, namely soft biometric characteristics, is used to provide the system with tracking capabilities. We demonstrate the feasibility of our approach by showing a real case application scenario.**

Keywords-video surveillance; multiple camera; global mapping; privacy-preserving; soft biometrics.

I. INTRODUCTION

With the fast increase of the number of video sensors employed in surveillance applications, several crucial challenges must be solved. On the one side, problems usually arise due to the large amount of data that makes overwhelming and no longer feasible human supervision. On the other side, this quantity of data generates important concerns about the privacy of people under surveillance.

In the first case, in addition to tiredness and loss of focus of attention an important problem to overcome is the loss of the *situation awareness* of the surveillance agent. That is to say, the confusion generated from monitoring an activity presented as two different and non-correlated views, especially if the areas to supervise are unknown [8]. In the second case, we would like to enforce the privacy of surveilled users without loosing, at the same time, the possibility of tracking a person as he/she moves across the surveilled areas.

A possible solution to the first challenge would be increasing the number of video supervisors. It is straightforward that such solution is not scalable since it requires large human efforts, thus presenting important economic drawbacks. Moreover, operators need accurate planning of time schedules to maximize their focus of attention, and site inspections are necessary to acquaint the agents with on site knowledge [9].

To face those limitations, many approaches have been explored that try to automate and ease the work of surveillance operators. In [8] a 3D environment is simulated where the images from a distributed camera network are projected on a reconstruction of surveilled areas. The reconstructed space aims at increasing scene understanding and people localization inside the scene. A similar approach was envisaged by [11] that added audio cues and time analysis for the events in a surveillance network, allowing a more precise localization of subjects within the video sensor network. Another approach employed in large-scale systems enforces situation awareness by linking cameras with the maps of the building. The system described in [10] displays camera locations within a map so that a spatial reference is always available. Still the agent has to make a mental effort to link images and camera locations on the map.

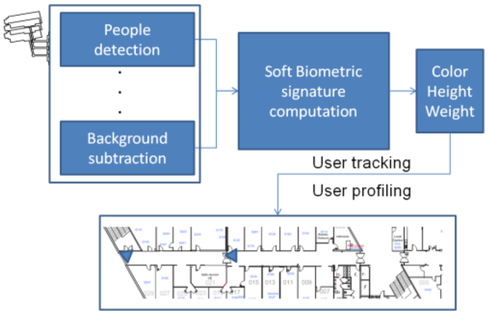

Figure 1. Scheme that summarizes the modules of our system. Images are analyzed in order to compute the soft biometric signature. Such signature is afterward used both to re-identify and to profile the surveilled person.

Another problematic aspect of video sensor networks are privacy concerns. Typical approaches rely on obfuscating the images coming from the cameras involved in the recording. According to the level of obfuscation a person may be partially (e.g. only the face) [12], or totally [13] hidden. While the partial obfuscation leaves clues useful to perform visual tracking (i.e. body shape and color information), it is clear that, even if coupled with automatic tracking, the total obfuscation creates further difficulties to the surveillance agents.

We seek to solve both issues by using a map that summarizes all the ongoing activities in a surveillance network (see Figure 1.). Additionally it provides a human compliant description of the physical appearance that makes feasible the re-identification task. The proposed framework provides an automatic solution for re-identifying people while respecting their privacy. We first establish a link between 3D spaces of a building recorded by multiple video sensors with its plan.

Secondly, we track and re-identify people moving around the building thanks to soft information on bodies. Finally, we display information concerning recorded people as gradual function of existing authorizations and/or requirements.

We perform single camera tracking that enables us to extract Soft Biometrics traits [7] that describes physical attributes of the person. We later use those attributes to re-identify the person in another camera. We do not make use of temporal clues, neither our cameras are overlapping. Thus, we only rely on the distance computed from soft biometric traits to re-identify the subject. Our design automatically deals with privacy issues: the representation of the user in the map is indeed completely anonymous. The tracking capability is maintained by the soft information provided as person's description.

In the remaining part of the paper, we introduce the mapping technique that creates the link between images and building plan in Section II. Then, in Section III we present the soft biometric signature that is used to re-identify and to describe people. In the last Section, we provide examples of our application scenario.

II. Linking video camera outputs and building map

The system we propose aims at generalizing and summarizing an entire video surveillance system in a global view that corresponds to the map of the covered areas (see Figure 1.).

Although the proposed approach best fits the scenario of indoor surveillance, it is generic enough to be applied to almost any building (e.g. undergrounds, airports). We require that a map of the area under surveillance is available (which is often the case for any building). Information from surveilled areas is exploited to create a correspondence between images and the building's map. Our method exploits the existing furnishings (e.g. doors), shared among different views as invariant features to create the mapping. Thanks to these elements, it is possible to inter-calibrate images from the camera with the map of the building.

A. Homography mapping

Obtaining the ground plane location of people in the scene means we should project the estimated positions for each frame on a single picture that represents the building map.

The invertible mapping from points in the image plane to points in the map plane is a planar projective transformation since it maps lines to lines (three collinear points in the image plan will still lay on the same line after projection and vice versa).

From an algebraic point of view, if we represent each two-coordinate point *(x, y)* as a homogeneous 3-vector ***p***=*(x, y, z)*, we can formulate a planar projective transformation as a linear mapping represented by a non-singular *3 x 3* matrix called homography matrix [14]:

$$(x' \ y' \ z')^t = H_{3x3}(x \ y \ z)^t$$

Therefore, once we have the homography matrix we can compute the ground plane location of a person on the map as a projection of the corresponding location of his/her feet on the image.

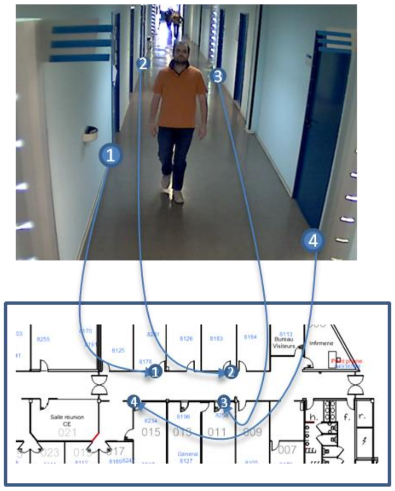

Figure 2. The image shows an example of point correspondences needed to compute the homography mapping which allows the transformation from the image space to the map plane. It is important to choose features that are shared between the two modalities.

We choose to solve the problem of estimation of 2D homography matrix by means of normalized Direct Linear Transformation (DLT) algorithm proposed by Hartley and Zisserman in [14]. The DLT algorithm exploits a set of corresponding points (***p′***↔***p***) to compute an estimate of the homography mapping between two planes. Since *H* is defined up to a scale factor, the transformation led by the 9 elements of *H* has 8 degrees of freedom. Since for each pair of corresponding points we have 2 linear equations in the *H* elements, it follows that if we have 4 corresponding points ***p′***↔***p*** then it exists only one solution *H* (if no more than 2 collinear points exist).

In our case, point correspondences are established using standard elements present both in the map and in the images. For our system (see Figure 2.), we have chosen the position of the doors on the floor to compute the homography matrix *H*. This is just an example of how to exploit information shared among the recorded scenes and the map of the building.

B. Single camera tracking

We exploit a tracking algorithm to trace the corresponding location of the surveilled subject on the map, and to compute the soft biometric signature. The adopted technique exploits a mix of algorithm for both detecting and tracking persons. In order to detect persons we use the well-known histogram of gradients approach from [17].

The feet position is estimated as the middle point of the bounding box previously extracted; then it is projected onto the building map thanks to the homography mapping. Finally, we track the position of the user in the map domain by using a Kalman filter approach. The tracking performs better in this domain since the constant velocity hypothesis is more suitable for such a scenario. In Figure 3. one can observe an example of tracked trajectories for two subjects walking in two different cameras.

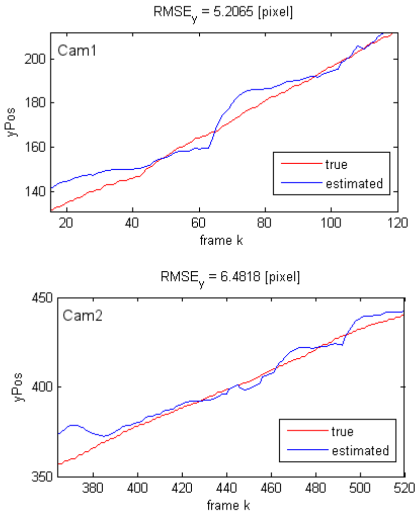

Figure 3. The two plots represent two examples of trajectories estimated from the ground plane location of the subject. A small value in *yPos* axis indicates that the subject is far away from the camera. The two tracks are compared to the ground truth (the straight line). It is noticeable how the uncertainty is generally bigger when the subject is far away from the camera.

III. SOFT BIOMETRICS BASED PEOPLE RE-IDENTIFICATION

Soft biometric traits are physical and behavioral human characteristics that help profiling people in pre-defined and human-compliant categories. They can be distinguished from the classical biometrics since they do not necessarily provide the same amount of information [7]. Among the soft biometrics, we can identify three main categories referred to body, face and accessories. The first one indicates traits that strongly relates to body appearance like height, weight, and gender. The second one refers to properties that describe facial details (e.g. hair color, ethnicity, and beard/mustache). The latter usually indicates the presence of additional characteristics that do not particularly belong to all but some people (e.g. tattoos, clothes color).

Hard biometric traits are intrinsically good for the identification task thanks to the greater complexity of their patterns (e.g. fingerprint and iris patterns). However, hard biometric traits represent a great challenge in case of video surveillance scenarios. Their acquisition is indeed problematic through camera normally employed in such situations. For this reason, body soft biometric traits provide an additional source of information about the identity of subjects in cases where the quality of video cameras does not match the requirements of classical biometrics (see Figure 4.).

In this paper we seek to provide a practical example of the possible uses of soft biometrics in a surveillance scenario. We propose the use of height, weight, and cloth colors to perform the quasi-immediate re-identification process when cameras do not overlap and when no temporal information is exploited.

In the following part, we will explore the three different components of our soft biometric signature and we will present our fusion scheme for the re-identification task.

A. Height

Stature is the first characteristic that determines one's physical aspect. Several methods exist to estimate the height of vertical objects in images the most known is presented in [1]. Some of these techniques were adapted to estimate people's height in videos or images [2]. Moreover, these algorithms can be divided in two classes depending on how they compute height. They can exploit information extracted from the captured scene [1], or they can use scene and camera knowledge, thus camera calibration [2].

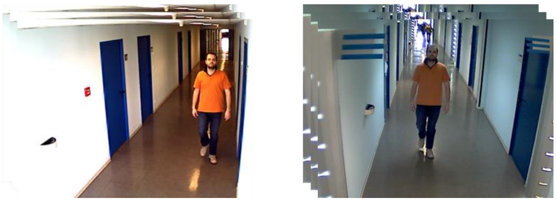

Figure 4. We show images from the sequences taken by two different cameras.

Since the cameras employed in surveillance systems are generally not calibrated, in our system we exploited the former approach. In our case, the knowledge of the scene is limited to the size of corridors' doors. No limitations exist to the application of this technique: it is indeed sufficient to find within the scenes a common target of know height to provide stature estimates**.** An example of height values measured from the images of our surveillance system is show in Figure 5.

In our case, doors' height does not change across Scene 1 and Scene 2. Therefore, we are able to compute the height of people as function of the doors' height (here used as measuring unit).

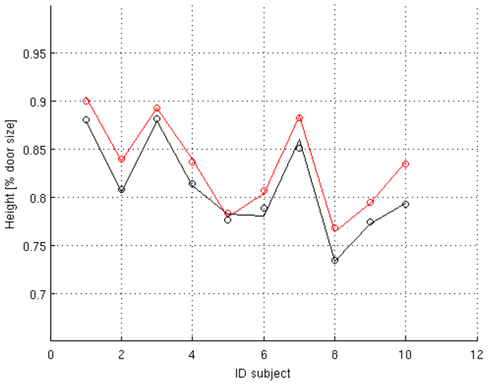

Figure 5. Values of height extracted from two different cameras. The values are reported as function of doors' height in the two different scenes. The graphs shows the height estimation results for two different sessions.

B. Weight

Weight is a recently exploited soft biometric trait. To the best of our knowledge, the only paper that involves weight [3] uses a scale to weigh users of a fingerprint recognition system for increasing the performance of the system by including this trait in the classification scheme. Another study [4] analyzed the possibility of estimating people's weight from images by using anthropometric measures. A series of anthropometric measures are extracted manually from images. Measures are applied to a linear regressor trained on large medical dataset to obtain the corresponding weight. In our case it was impossible to extract a set of anthropometric measures since our scenario involves uncalibrated cameras. Then, to provide a measure related to weight, we resorted to the area occupied by the silhouette of the person w.r.t. the distance of the person from the camera. The extraction of the foreground was performed with the algorithm presented in [6].

The measure of area provided is given as a ratio over the entire bounding box extracted with the people detector module. Together with the distance of the subject from the camera, and by filtering the results obtained in a window of several frames, we can guarantee enough independence of such result from the capturing camera.

C. Clothes color

Using colors for tracking people across a camera network can be challenging. As colored surfaces are exposed to different condition of illumination, the perceived color can vary substantially (see Figure 4.).

The probabilistic color histogram (PCH) presented in [5] is used to provide a color descriptor invariant to illumination conditions. Images are converted into Lab color space and the pixel further elaborated using a fuzzy K-nearest neighbor into the 11 culture colors classes (*red, orange, black, pink, white, gray, purple, brown, blue, yellow, and green*).

Consequently, for each given pixel we obtain its classification score for each given color out of the 11. The descriptors are in the form of a probabilistic color histogram (PCH) that represents the probability for a given pixel to be represented by one of the 11 classes previously mentioned.

An example of a probabilistic color histogram is shown in Figure 6.

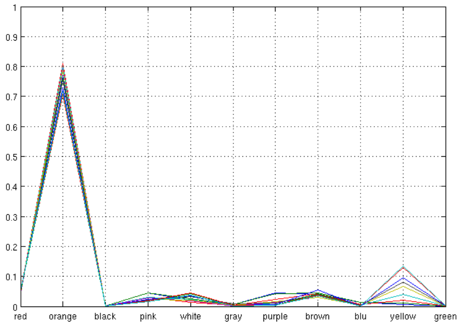

Figure 6. Considering pixel Lab components, the probabilistic color histogram measure the probability of belonging to a particular color class. In the figure the PCH of consecutive frames are shown. In this case the correlation among frames is evident.

D. Bag of body soft biometrics for re-identification

To extract information that describes each user (height, weight, and color) a foreground and background segmentation algorithm is necessary. We employed the commonly available system provided by [6]. The silhouette obtained, is used to estimate the occupied area (for weight estimation), and to separate the silhouette color from the background to compute the color descriptors.

Consequently to the features extraction, an approach similar to the one presented in [7] is exploited. In our case the bag of soft biometrics is limited to characteristics of physical body appearance.

Then, the results are fused together as follows:

$$D_{total} = \alpha\, D_{color} + \beta\, D_{height} + \gamma\, D_{weight} \qquad (1)$$

where each D is an Euclidean distance measure of the corresponding feature and α, β, and γ are weighting parameters chosen to provide more importance to the soft biometric trait which provides higher entropy than the others. The values chosen are 0.6, 0.2, and 0.2 respectively so that color (which is more discriminating) will weigh more.

To validate our idea we exploited a set up similar to the one presented in [16]. Then, we recorded video sequences of 10 people walking in a corridor environment. Challenges are presented by the varying illumination conditions and the compression applied to the videos (at the source). The videos recorded were analyzed with the processing steps already presented. Results of such fusion lead to the confusion matrix in Figure 7. All the subjects were correctly re-identified as moving from one camera to the other.

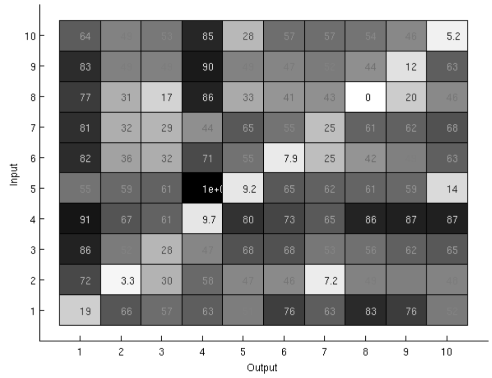

Figure 7. Confusion matrix representing the re-identification output of our system. The Input being the signature computed at Camera1, and Output being the outcome search in Camera2. The minimum distance subject along the rows gives the corresponding person among the two cameras.

IV. Displaying appropriate information

Video surveillance is a technology that provides many benefits and, at the same time, generates many concerns because users' privacy is generally involved. For this reason, in the past years, a common solution was to record images/videos without employing surveillance agents. This guarantees that, if needed, the videos could be available to law enforcement agencies. In other cases, the monitoring of qualified agents is required, thus guaranteeing their prompt intervention in case of emergency. Problems may arise as the number of camera grows because the number of required agents increases as well [9].

The two presented are objective limitations to the use of video surveillance systems. In the first one, the limit is the possibility of having only passive surveillance, loosing the ability of a prompt security intervention. In the second case, the limitation lies in the large number of surveillance agents needed to cope with the number of surveilled cameras.

In order to tackle both these problems we propose a system that is able to summarize a surveillance network by providing to the agent a digest of the scenes recorded. An example of such system can be observed in Figure 8. .

Each person in the video surveillance system is represented as a two-color square. The colors of the square represent the upper and lower part of the clothes (i.e. torso/legs) obtained from the soft biometric signature computation. Although the specific design of PCH descriptor provide us with slight intrinsic robustness to illumination variations, we compute the final color components used in the user interface as follows:

$$\mathrm{C}_{[\mathrm{torso|legs}]} = NPCH_{[\mathrm{torso|legs}]}\, C_{\mathrm{matrix}} \qquad (1)$$

where NPCH is the normalized probabilistic color histogram (computed as PCH/‖PCH‖) and $\mathrm{C}_{\mathrm{matrix}}$ is the color components matrix for the 11 culture colors chosen in our setup. Each of the RGB components of the culture colors where chosen from the values established by the color naming experiments performed by [15].

Our schema does not provide to the security agent images of the live recordings; this allows tracking people moving around the surveilled camera network, while ensuring their privacy. The description provided by our human compliant signature can be used to identify a person provided also their position. Furthermore, videos can be recorded so that law enforcement agents could access them if needed.

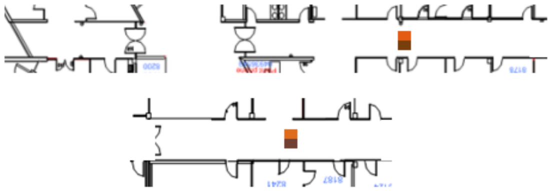

Figure 8. Example of the output of our system. It shows portion of the same corridor recorded by two cameras in two different moments and under different light conditions. The color of the top of the square matches the color of the torso, while the bottom part matches the color of the legs.

V. Conclusion

We presented a new approach to video surveillance that aims at increasing situation awareness of surveillance operators while keeping an eye on both the re-identification and the privacy of the surveilled users.

Our system combines an algorithm for single camera tracking, a re-identification module based on soft-biometric signature, and a visualization interface that preserves privacy of users. We exploit geometrical correspondences between the scene and the map to compute a homography that links people with their current location in the scene. Small colored rectangles depict the users inside the map of the building. We exploit body soft biometrics (height, weight, and color) to create a signature that allows re-identification of people as they move through the camera network. Moreover, we use the soft biometric signature to enforce the situation awareness of the surveillance agent by creating a correspondence between the colors of persons' representations as they move in the map.

Our system is able to summarize the information from the scene; it provides a good solution in terms of privacy preservation and security management of buildings. In order to provide an additional level of privacy preservation a similar system could provide ways of filtering displayed information to the user according to a given level of rights. In this way, users with higher permissions can easily access to the entire content of the surveillance system, while low ranks allow only the supervision of the essential information provided by the system. Another possible future upgrade, would link additional soft biometric traits (like hair color, gender, and ethnicity) with the shape of the person representation in order to provide additional clues for the visual tracking.

References

[1] A. Criminisi, I. Reid and A. Zisserman , Single view metrology, International Journal of Computer Vision , Springer, 2000, pp. 123–148

[2] S.H. Lee and J.S. Choi , A Single-View Based Framework for Robust Estimation of Height and Position of Moving People, Advances in Image and Video Technology, Springer, pp. 562–574

[3] H. Ailisto, E. Vildjiounaite, M. Lindholm, S.M. Makela and J. Peltola , Soft biometrics-combining body weight and fat measurements with fingerprint biometrics, Pattern Recognition Letters, 27.5, 2006

[4] C. Velardo and J.-L. Dugelay, Weight estimation from visual body appearance, Biometrics: Theory, Applications and Systems, 2010

[5] A. D'Angelo and J.-L. Dugelay, People re-identification in camera networks based on probabilistic color histograms, Visual Information Processing and Communication, SPIE Electronic Imaging, 2011

[6] L. Li, W. Huang, I.Y.H. Gu and Q. Tian, Foreground object detection from videos containing complex background, ACM international conference on Multimedia, 2003, pp. 2-10

[7] A. Dantcheva, C. Velardo, A. D'angelo and J.-L. Dugelay, Bag of soft biometrics for person identification : New trends and challenges, Mutimedia Tools and Applications, Springer, October 2010

[8] S. Fleck, F. Busch, P. Biber, and W. Straber, 3D surveillance - a distributed network of smart camers for real-time tracking and its visualization in 3D. CVPRW, 2006

[9] M. M. Trivedi, T. L. Gandhi, and K. S. Huang, Distributed interactive video arrays for event capture and enhanced situational awareness, IEEE Intelligent Systems, Special Issue on Homeland Security, 2005

[10] Y. Tian, L. Brown, A. Hampapur, M. Lu, A. Senior, and C. Shu, IBM smart surveillance system (S3): event based video surveillance system with an open and extensible framework, Machine Vision and Applications 2008, v. 19-5, pp. 315-327

[11] M. Baklouti, M. Chamfrault, M. Boufarguine, V. Guitteny, Virtu4D : A dynamic audio-video virtual representation for surveillance systems, 3rd International Conference on Signals Circuits and Systems 2009, v. 6-8, pp.1-6, November 2009

[12] F. Dufaux, T. Ebrahimi, Scrambling for Video Surveillance with Privacy, Computer Vision and Pattern Recognition Workshop, 2006. vol.160, pp. 17-22 June 2006

[13] M. Upmanyu, A.M. Namboodiri, K. Srinathan,C.V. Jawahar, Efficient privacy preserving video surveillance, ICCV 2009, pp.1639-1646, October 2009

[14] R.I. Hartley, and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2000

[15] Experiments on color naming - http://xkcd.com/color/rgb/

[16] R. Chellappa, P. Turaga, Recent Advances in Age and Height Estimation from Still Images and Video, Automatic Face and Gesture Recognition, March 2011

[17] N. Dalal, and B. Triggs, Histograms of oriented gradients for human detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1, 2005