Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 27 (2006) 1809–1815
Pattern Recognition Letters
www.elsevier.com/locate/patrec

A system for automatic face obscuration for privacy purposes

Rita Cucchiara [a], Andrea Prati [b,*], Roberto Vezzani [a]

[a] *Dipartimento di Ingegneria dell'Informazione, University of Modena and Reggio Emilia, Via Vignolese 905, 41100 Modena, Italy*
[b] *Dipartimento di Scienze e Metodi dell'Ingegneria, University of Modena and Reggio Emilia, Via Allegri 13, 42100 Reggio Emilia, Italy*

Available online 3 May 2006

Abstract

This work proposes a method for automatic face obscuration capable of protecting people's identity. Since face detection heavily benefits from the possibility to exploit tracking, multi-camera people tracking has been integrated with a face detector based on colour clustering and Hough transform. Moreover, the multiple viewpoints provided by multiple cameras are exploited in order to always obtain a good-quality image of the face. The identity of people in different views is kept consistent by means of a geometrical, uncalibrated approach based on homographies. Experimental results show the accuracy of the proposed approach.

Keywords: Consistent labelling; People tracking; Face detection; Multi-camera tracking

1. Introduction

Recent events all over the world have contributed to increasing the demand for security of the citizens. As a consequence, both industrial companies and public entities have invested a great deal of time and many resources in security-related problems. However, one of the fundamental rights of citizens is the protection of their privacy. Most of the western countries have a set of more or less restrictive laws to assure that their citizens' privacy is respected.

For this reason, there is an emergent need for (semi-) automatic tools for protecting people's identity, especially in public video surveillance. A possible solution is to use PIR (passive infrared) sensors to detect (and track) people anonymously, while cameras are used in public areas to obtain people's identities, but only when necessary. An alternative solution relies on the capability given by computer vision algorithms to recognize humans and, specifically, their faces. Artificially obscuring faces is an effective way to protect identities and, at the same time, save the face images for further, authorized accesses for security purposes. Unfortunately, even though face detection techniques are now mature, it can be hard to obtain a frontal view of the face in complex environments. To help in this task, multiple cameras can be used in order to obtain different views and, potentially, at least a frontal view of the face. In addition, a multi-camera vision system enables covering wider areas and multiple viewpoints provide an effective solution to the problem of occlusions in cluttered scenes. Merging data provided by multiple cameras can lead to several problems. The main problem arises when objects move from the field of view of a camera to that of another camera. In this case, the objects' identity must be preserved, in order to analyze their behaviour over the whole scene. In literature, this process is known as *consistent labelling* and becomes challenging when cameras cannot be manually calibrated.

In this paper, we report on a novel approach for consistent labelling with automatic learning of homographic transformation between the ground planes of overlapped cameras. Moreover, the techniques adopted to track multiple people and provide people tracking and face detection with the purpose of obscuring the people's faces are described. In particular, the module for people tracking on multiple cameras and the module for face detection are closely integrated in order to improve face detection

* Corresponding author.
E-mail addresses: cucchiara.rita@unimore.it (R. Cucchiara), prati.andrea@unimore.it (A. Prati), vezzani.roberto@unimore.it (R. Vezzani).

0167-8655/$ - see front matter
doi:10.1016/j.patrec.2006.02.018

by exploiting tracking and to provide different views of the same person (provided by different cameras). For this purpose, the same label among different views must be assigned to each instance of the same person, that is consistent labels must be guaranteed. A preliminary version of this work was published in (Cucchiara et al., 2005). In spite of the previous considerations, the next two sections will describe, respectively, the method used to establish consistent labelling and the approach used for face detection. Finally, Section 4 describes the experimental results.

2. Consistent labelling

The problem of consistent labelling has been addressed in the literature in two main ways. The first relies only on the appearance of the objects, the second on the geometrical relationship between overlapped cameras. In the appearance based approaches the matching is essentially based on the colour of the tracks, by using invariants to light changes and texture features, and clustering that is based on mean shift (Li et al., 2002) or matching of colour histograms (Krumm et al., 2000). In the case of cameras with non-overlapped views, this is the only possible solution. However, if the camera views are partially overlapped, using merely the object's appearance is not a successful strategy, since the appearance (in particular, the colour) can be reproduced very differently with different cameras and under different illumination conditions. As a consequence, other works in the literature are based on geometrical constraints. Geometry-based approaches can be further subdivided into calibrated (Mittal and Davis, 2001; Yue et al., 2004) and uncalibrated (Khan and Shah, 2003) approaches. The approach proposed by Khan and Shah (2003) is based on the computation of the so-called Edges of Field of View (hereinafter referred to EoFoV), i.e., the lines delimiting the field of view of each camera and thus defining the overlapped regions.

Similar to the proposal of (Khan and Shah, 2003), let us suppose that the system is composed of a set $\boldsymbol{C} = \{C^1, C^2, \ldots, C^n\}$ of n cameras, with each camera C^i overlapped with at least another camera C^j. Let us call 3DFOV lines $L^{i,s}$ the projection of the limits of the field of view (FoV) of camera C^i on the ground plane ($z = 0$), corresponding to the intersection between the ground plane and the rectangular pyramid with its vertex at the camera optical center (the camera view frustum); s indicates the equation of the line on the image plane. In particular, four of them, L^{i,s_h}, $h = 1, \ldots, 4$ could be computed, with s_h corresponding to the image borders $x = 0$, $x = x_{\max}$, $y = 0$, and $y = y_{\max}$. They could be visible also by another camera; in such a situation, we define as EoFoV$L_j^{i,s}$ the 3DFOV line corresponding to s of camera C^i seen by camera C^j. EoFoV$L_j^{i,s}$ divides the image on camera C^j into two half-planes, one overlapped with camera C^i and the other one disjointed. The intersection of the overlapped semi-planes defined by the EoFoV lines from camera C^i to camera C^j generates the overlapping area Z_j^i.

The EoFoV lines are created with a training procedure. A single person moves freely in the scene, with the minimum requirements to pass through at least two points of each limit of the FoV of two overlapped cameras. Let us call O_k^i the object segmented and tracked with label k in the camera C^i and SP_k^i the point of contact with the ground plane (hereinafter referred to as *support point*). The support point can easily be computed as the middle point of the bottom of the bounding box of the blob.

Given the constraint to have a single moving person in the training video, problems of consistent labelling do not occur. Thus, when the object is detected also by camera C^j and tracked with label p, it is directly associated with O_k^i. Therefore, at this moment (known as the moment of "camera handover"), the support point SP_k^i can be associated with SP_p^j (if it is visible). In this case the point SP_k^i lies on the EoFoV line $L_i^{j,s}$ for camera C^i. The equation of each line $L_i^{j,s}$ is computed by collecting a set of coordinates of the support point SP_k^i detected at the camera handover and exploiting a least square optimization (Fig. 1(a)). However, there are cases where, at the moment of camera handover, the detected parts of the person do not lie on the ground plane, as in Fig. 1(b), where the head is detected. Thus, matching the point of a head in this camera with the SP in the other camera is incorrect and causes an erroneous EoFoV computation.

To solve this problem, we modified the approach proposed by Khan and Shah (2003) by delaying the computation of the EoFoV lines to the moment in which the object completely enters the scene of the new camera (see Fig. 1(c)). This can lead to a displacement of the line with respect to the actual limit of the image, but it assures correct matching of the position of the feet in the two views. As a consequence, the actual FoV lines are neither coincident nor parallel to the image border. Since, for our approach to the consistent labelling, the choice of the line used to create the EoFoV is completely arbitrary, it does not impact on the result of the calibration. Obviously, the more the selected lines are closer to the centre, the more imprecise the homography is.

The approach proposed by Khan and Shah (2003) establishes the consistent labelling only in the exact moment of the camera handover from C^i to C^j. This approach has two main limitations: if two or more objects cross simultaneously (Fig. 2) an incorrect labelling can be established; if they are merged from the view of C^j at camera handover, but then they separate, the consistent labelling with the labels of C^i cannot be recovered (Fig. 3).

We propose to overcome these problems by means of homography, thus extending the matching search to the whole overlap zone of the field of view. For two overlapped cameras C^i and C^j, the training procedure computes the overlapping areas Z_j^i and Z_i^j. The four corners of each overlapping area Z_j^i and Z_i^j define the sets of points $P_j^i = \{p_1^{i,j}, p_2^{i,j}, p_3^{i,j}, p_4^{i,j}\}$ and $P_i^j = \{p_1^{j,i}, p_2^{j,i}, p_3^{j,i}, p_4^{j,i}\}$ where the subscripts indicate corresponding points in the two cameras (see Fig. 1(c)). These four associations between points

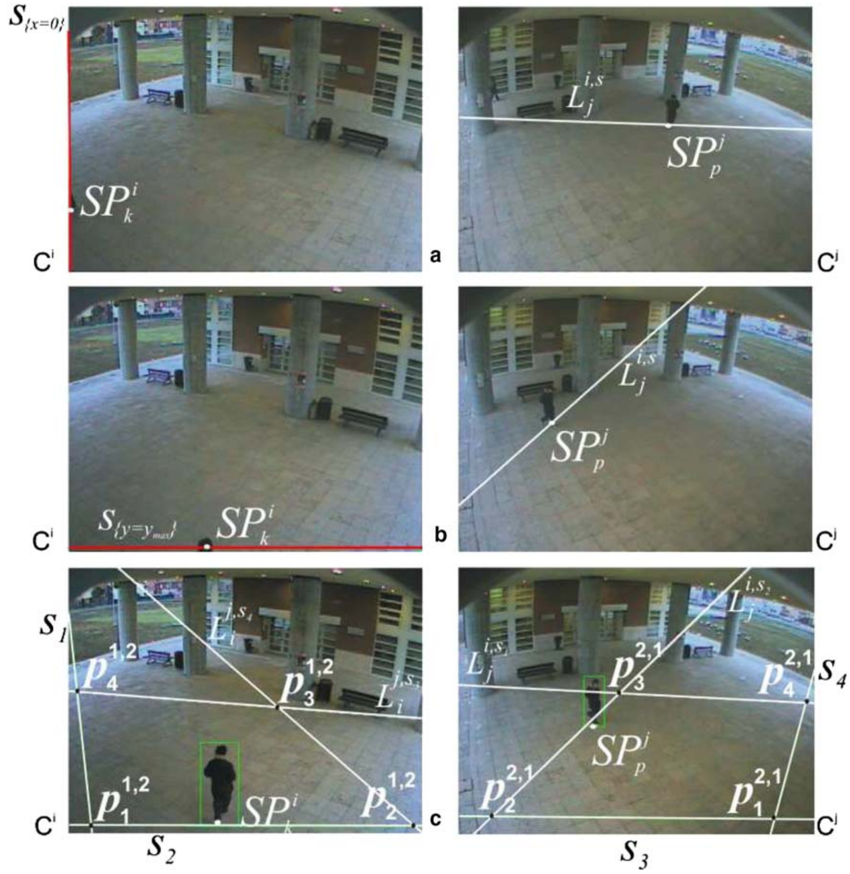

Fig. 1. Examples of EoFoV computation.

Fig. 2. Examples of simultaneous transition.

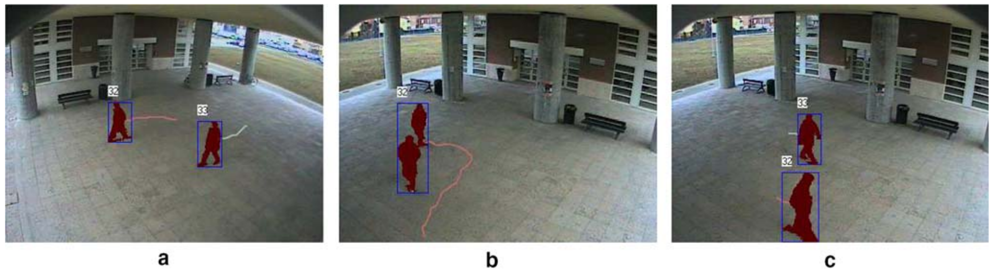

Fig. 3. Examples of merged transition: (a) C^2 at frame 1250, (b) C^1 at frame 1250 and (c) C^1 at frame 1260.

of camera C^i and points of camera C^j on the same plane $z = 0$ are sufficient to compute the homography matrix H_j^i from camera C^i to camera C^j. Obviously, the matrix H_i^j can easily be obtained with the equation $H_i^j = (H_j^i)^{-1}$.

Each time a new object O_k^i is detected by camera C^i inside the overlapping area (not only at the moment of camera handover), its support point SP_k^i is projected in C^j by means of the homographic transformation. Calling $(x_{\mathrm{SP}_k^i}, y_{\mathrm{SP}_k^i})$ the coordinates of the support point SP_k^i, we can write the projected point in homogeneous coordinates $[a,b,c]^{\mathrm{T}} = H_j^i[x_{\mathrm{SP}_k^i}, y_{\mathrm{SP}_k^i}, 1]$. The projected point $\widetilde{\mathrm{SP}_k^j}$ corresponds on the image plane of C^j to the projective coordinates $\tilde{x}^j = a/c$ and $\tilde{y}^j = b/c$. These coordinates could not correspond to the support point of an actual object. For the match with object O_k^i, we select the object in C^j whose support point is at the minimum distance on the 2D plane from these coordinates

$$O_k^i \leftrightarrow O_p^j \bigg| p = \arg\min_q D\left(\widetilde{\mathrm{SP}_k^j}, \mathrm{SP}_q^j\right) \quad \forall q \in \boldsymbol{O}^j \tag{1}$$

where $D(\cdot)$ denotes the Euclidean distance and $\boldsymbol{O}^j$ is the set of objects detected by C^j. The results achieved with this approach in the two above-mentioned cases are shown in Figs. 2 and 3(c), respectively, where the correct label assignment is obtained.

In conclusion, it is worth noting that the proposed algorithm is almost independent of the method used for object detection and tracking for single cameras. In our case, we used the SAKBOT (statistical and knowledge-based object tracker) system proposed by Cucchiara et al. (2003) and the tracking procedure described by Cucchiara et al. (2004).

3. Face detection through object tracking

Once moving people are detected and tracked by the whole multi-camera system, we can collect a set of different views of the same person. Among these views, we can assume that at least one can provide a frontal view of the person for easier face detection (for the twofold purpose of recognition and face obscuration).

Face detection is a widely explored research area in computer vision. Two recent surveys, (Yang et al., 2002) and (Hjelm and Low, 2001), have assembled a large number of proposals regarding face detection. Most of them are based on skin colour detection (Jones and Rehg, 2002) followed by face candidate validation achieved by exploiting geometrical and topological constraints.

Unfortunately, most of the colour-based approaches are computationally very expensive and it is impossible to perform accurate face detection at every frame in a real-time video surveillance application. To solve this problem, the face detection can be performed only when a new person enters the scene and then a face tracking can be adopted. This approach requires reliable people tracking as fundamental step.

Our method exploits and improves the best ideas proposed by Birchfield (1998) and Maio and Maltoni (2000). The former uses both colour and gradient information but the search of the head is limited to a neighbourhood of a predicted position. Unfortunately, this solution requires a high frame rate to make reliable predictions. Maio and Maltoni (2000), on the other hand, adopt a solution based on the elliptical Hough transform; unlike the previous one, this solution does not require any tracking or prediction, as the processing of each frame is standalone. A face colour histogram must be available as a model.

To this end, a supervised learning phase is performed to compute a histogram H of skin and hair colours. We have collected a set of about 400 heads obtained through manual segmentation of training videos. Regular colour histograms are computed for each of these samples and integrated in a global histogram to obtain the reference model. Heads with different rotations, sizes and light conditions have been included in the training set to make the model as general as possible. To reduce the size of the stored histogram and to speed up the subsequent comparisons, we adopted a compressed colour space based on the three axes B–G, G–R, and $B+G+R$ (Swain and Ballard, 1991).

Thus, for each tracked object O_j, two different Hough transforms are computed: one gradient-based T_{g} and one colour-based T_{c}. The points belonging to the edges of the track (obtained with Sobel edge detectors) vote for the first transform according to the gradient value. The selection of the voted pixels is done by moving on the image in the same gradient direction with a distance obtained from the estimated head size (see Fig. 4). Calling α the angle of the gradient of the point (x, y) with respect to the horizontal axis, a and b the horizontal and the vertical half-sizes of the ellipse, respectively, we can obtain the coordinates of the two candidate centers of the face $(\mathrm{FC}_1, \mathrm{FC}_2)$

$$\begin{aligned} &\Delta x = \frac{a^2}{\sqrt{a^2 + b^2 \cdot tg^2\alpha}} \quad \Delta y = \frac{b^2}{a^2} \cdot tg\alpha \cdot \Delta x \\ &\mathrm{FC}_1 = (x - \Delta x, y - \Delta y) \quad \mathrm{FC}_2 = (x + \Delta x, y + \Delta y) \end{aligned} \tag{2}$$

Similarly, a point of the object votes for the colour-based transform if its colour has a non-zero value on the histogram H. In this case, it votes for all the points inside an ellipse having the same size of the head and the current pixel as the centre, and the rate is proportional to the

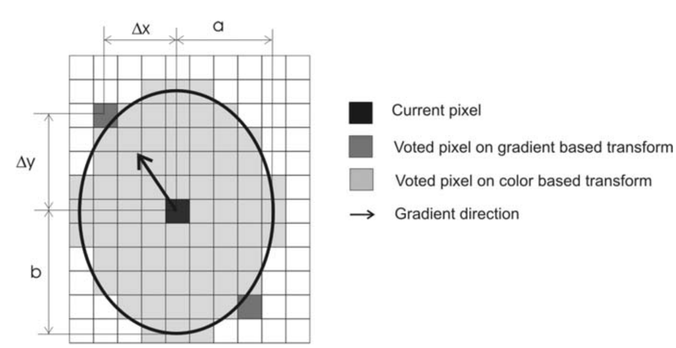

Fig. 4. Computation of the elliptical Hough transforms.

a b

Fig. 5. Examples of face obscuration (a) and avoidable face obscuration (b).

Table 1
Experimental results using OpenCV face detector on faces of different sizes

	TN	FS > 54 px		48 px < FS < 54 px		40 px < FS < 48 px		FS < 40 px	
		N_1	R_1 (%)	N_2	R_2 (%)	N_3	R_3 (%)	N_4	R_4 (%)
FC_1	1005	585	100	100	100	120	60	200	0
FC_2	750	416	100	100	91	134	42.537	100	0
NFC_1	1032	432	100	99	67.68	101	28.713	400	0
NFC_2	727	360	100	115	68.7	114	20.175	138	0

C_n: camera n; TN: total number of frames of the video; FS: face size (the face is considered about FS × FS pixels); N_n: number of frames of the video in range n; R_n: number of correct detections in range n.

model histogram value corresponding to the colour of the pixel. Thereafter, the two transforms are normalized and multiplied pixel-by-pixel to obtain a single map that contains both colour and gradient information. The point with the higher value is chosen as the centre of the head of the object O_j. Once the face is detected and tracked, the head can be obscured, as shown in Fig. 5(a).

It is worth noting that the size of the face is a crucial constraint for both detection and recognition. Typical algorithms of face detection can only be used with a sufficient resolution, in other words, when the face is acquired at a reasonable size. As proof of the concept, we tested one of the best assessed algorithms of face detection, namely the Viola–Jones approach (Viola and Jones, 2001). This method proposes the adoption of the Haar transform to create patterns of interest and the AdaBoost classifier to identify pixel patterns that can be considered as "faces". Table 1 shows an example of results with two different webcams (C_1 and C_2) taken at four different resolutions with frontal (F) and non-frontal (NF) (±15) poses. Ten versions of the sixteen situations have been replicated. Results show that in our experiments, using the algorithm in OpenCV library (http://www.intel.com/research/mrl/research/opencv/, 2005), when the face is larger than 54 × 54 pixels, face detection is always corrected. The correctness is acceptable with more than 48 × 48 pixels. For smaller sizes, the correctness is reduced. No face detection is possible for faces smaller than 40 × 40 pixels.

Regarding this last consideration, for the UK regulations on "privacy and forensic use of video material in CCTV systems" (Aldrige and Gilbert, 1995), a frame

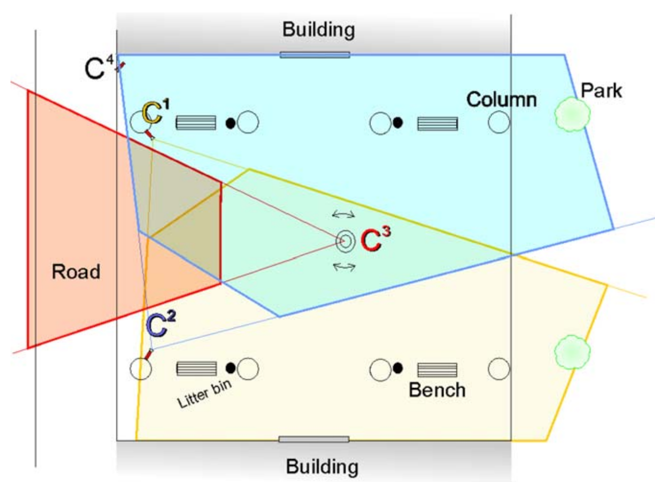

Fig. 6. Sketch of the test bed.

is suitable for recognition[1] and identification[2] if the head's height is at least 39 and 93.5 pixels, respectively. Only then is the above mentioned size-limit of 40 pixels justified.

However, if the face is too small, the person's identity is already protected by the low image resolution, as shown in Fig. 5(b).

[1] Recognition means that the viewer can identify that the person seen is the same, having seen that person before.

[2] Identification assumes that picture quality and details are sufficient to enable the identification of a subject beyond reasonable doubt.

a b c d

Fig. 7. Some snapshots of the system output after consistent labelling: (a) C^1 at frame #783, (b) C^2 at frame #783, (c) C^1 at frame #1080 and (d) C^2 at frame #1080.

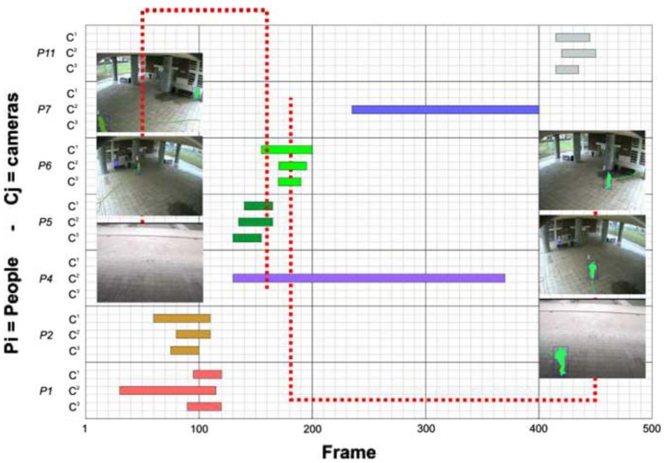

Fig. 8. Visibility and labels (indicated with the colour of the bars) of the tracks in a test sequence. (For interpretation of the figure in colour, the reader is referred to the web version of this article.)

4. Experimental results

To test our algorithms we created a test bed on our campus, installing four partially overlapped cameras (three fixed and one Pan–Tilt–Zoom—PTZ), as shown in Fig. 6, in a zone through which many people pass: there are some benches and the light conditions are typical of an outdoor environment.

The consistent labelling algorithm has been tested extensively with partially overlapped cameras. Some snapshots of the system output (in non-trivial conditions) after the consistent labelling assignment are illustrated in Fig. 7.

The track graph in Fig. 8 reports, for each person P_i, the slot of time (in frames) in which it is visible by the three cameras (C^1, C^2, and C^3) of our real setup. The colour of the bars corresponds to the identifier assigned by the consistent labelling algorithm.

We have tested the system also in the presence of simultaneous transitions of more than one person at a time (sync. trans.) and in the presence of transitions where two people are merged (merged trans.) in a single track during camera handoff and split far from the EoFoV. Table 2 shows the results obtained: the number of camera transitions correctly identified (in which consistent labelling is established) and the number of wrong correspondences are shown in the last two columns of the table. It is evident that the system is extremely accurate. The incorrect matches are mainly due to errors in the lower modules, i.e., in the segmentation and single camera tracking algorithms.

Table 3 shows, instead, some results of the face detection with a people tracking algorithm. Unlike the results shown in Table 1, face detection is not carried out at frame level, but the head is detected initially and then tracked. In this case, we obtain two important results: the face is detected at lower resolution (less then 40×40 pixels) and even in different poses (not only with a frontal view).

Table 2
Experimental results

Video	Sync. trans.	Merged trans.	No. of frames	No. of trans.	Correct	Incorrect
V1	No	No	8500	41	39	2
V2	No	No	3000	5	5	0
V3	Yes	No	1800	14	13	1
V4	Yes	Yes	2000	7	6	1
V5	Yes	Yes	500	2	2	0

Table 3
Performance of the face detection with people tracking module

Video	No. of frames	% Recogn.	Frontal view	Lateral view	Lateral horizontal view	Top view	Mean face size
V3	328	100	104	107	0	117	31×39
V4	440	99	112	162	166	0	25×31

5. Conclusions

The aim of this work was to propose a (semi-)automatic solution to the problem of protecting people's identities in public video surveillance. The proposed solution only relies on computer vision to detect and track people and to automatically obscure their faces. The use of other types of sensor may be either unaffordable or too expensive. Moreover, computer vision is able to extract faces and, at the same time, keep them for further uses. The proposed method also uses multiple cameras to control wider areas and to provide several views of people's faces.

Experimental results demonstrate that both the multi-camera tracking and the face detection achieve high accuracy.

Acknowledgements

This work was supported by the project LAICA (Laboratorio di Ambient Intelligence per una Città Amica), funded by the Regione Emilia-Romagna, Italy.

References

Aldrige, J., Gilbert, C., 1995. Performance Testing CCTV Perimeter Surveillance Systems. Police Scientific Development Branch (PSDB) Publication, United Kingdom Home Office. Available from: <http://scienceandresearch.homeoffice.gov.uk/hosdb/pdfs/14-95perftestingcctv.pdf>, website accessed: 09 January 2006 (14).

Birchfield, S., 1998. Elliptical Head Tracking Using Intensity Gradients and Color Histograms. In: 1998 Conf. on Computer Vision and Pattern Recognit. (CVPR '98), June 23–25, 1998, Santa Barbara, CA, USA. IEEE Computer Society, pp. 232–237.

Cucchiara, R., Grana, C., Piccardi, M., Prati, A., 2003. Detecting moving objects ghosts and shadows in video streams. IEEE Trans. Pattern Anal. Mach. Intell. 25 (10), 1337–1342.

Cucchiara, R., Grana, C., Tardini, G., Vezzani, R., 2004. Probabilistic people tracking for occlusion handling. In: Proc. of 17th Int. Conf. on Pattern Recognit. (ICPR 2004), 23–26 August 2004, Cambridge, UK, vol. 1. IEEE Computer Society, pp. 132–135.

Cucchiara, R., Prati, A., Vezzani, R., 2005. Ambient Intelligence for Security in Public Parks: The LAICA Project. In: Proc. of IEE Int. Symp. on Imaging for Crime Detection and Prevention, London, UK, 7–8 June, 2005, pp. 139–144.

Hjelm, E., Low, B., 2001. Face detection: A survey. Comput. Vision Image Understand. 83 (3), 236–274.

Jones, M., Rehg, J., 2002. Statistical color models with application to skin detection. Int. J. Comput. Vision 46, 81–96.

Khan, S., Shah, M., 2003. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. IEEE Trans. Pattern Anal. Mach. Intell. 25 (10), 1355–1360.

Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., Shafer, S., 2000. Multi-camera multi-person tracking for EasyLiving. In: Proc. IEEE Int. Workshop on Visual Surveillance (VS '00), July 1, 2000, Dublin, Ireland. IEEE Computer Society, pp. 3–10.

Li, J., Chua, C., Ho, Y., 2002. Color based multiple people tracking. In: Proc. IEEE Int. Conf. on Control, Automation, Robotics and Vision, December 2–5, 2002, Marina Mandarin, Singapore, vol. 1. IEEE Press, pp. 309–314.

Maio, D., Maltoni, D., 2000. Real-time face location on gray-scale static images. Pattern Recognit. 33 (9), 1525–1539.

Mittal, A., Davis, L., 2001. Unified multi-camera detection and tracking using region-matching. In: Proc. IEEE Workshop on Multi-Object Tracking, WOMOT'01, July 8, 2001, Vancouver, BC, Canada. IEEE Computer Society, pp. 3–10.

Open Source computer vision (OpenCV) library, Intel Corporation, 2005. Available from: <http://www.intel.com/research/mrl/research/opencv/>, website accessed: 09 January 2006.

Swain, M., Ballard, D., 1991. Color indexing. Int. J. Comput. Vision 1 (7), 11–32.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: Proc. of 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognit. (CVPR 2001), with CD-ROM, 8–14 December 2001, Kauai, HI, USA, vol. 1. IEEE Computer Society, pp. 511–518.

Yang, M., Kriegman, D., Ahuja, N., 2002. Detecting faces in images: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 24 (1), 34–58.

Yue, Z., Zhou, S., Chellappa, R., 2004. Robust two-camera tracking using homography. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, May 17–21, 2004, Montreal, Quebec, Canada, vol. 3. IEEE Signal Processing Society, pp. 1–4.