Ambient Intelligence for Security in Public Parks: the LAICA Project

R. Cucchiara, A. Prati, R. Vezzani
University of Modena and Reggio Emilia, Italy

ABSTRACT

In this paper, we address the exploitation of computer vision techniques to develop multimedia services and automatic monitoring systems related to the security and the privacy in public areas. The research is part of a two-year Italian project called *LAICA*, intended to provide advanced services for citizens and public officers. Citizens want fast and friendly web access to public places, to see the environment in real-time without violating the privacy laws. Public officers and policy centres want a fast and reactive monitoring system, capable to automatically detect dangerous situations, given the huge amount of cameras that can not be monitored simultaneously by human operators. In this work, we describe the project and the defined methodologies in multi-camera video mosaicing, people tracking and consistent labelling, and access to processed data with face obscuration.

1. INTRODUCTION

The recent events of terrorist attacks all over the world have contributed in increasing the request for security of the citizens. As a consequence, both industrial companies and public entities have invested much time and resources in security-related problems.

This paper reports on a research, part of a project called LAICA (Laboratorio di Ambient Intelligence per una Città Amica – Laboratory of Ambient Intelligence for a Friendly City) funded by the Regione Emilia-Romagna (Italy) and in collaboration with the municipality of Reggio Emilia, Italy (2004-2006), and several Italian universities and industrial companies. This multi-disciplinary project brings together the academic expertises and the industrial knowledge into several fields, from the low-power sensor networks, to the computer vision, to the middleware and mobile agents, to the communication. The objective of the project is the study and development of advanced services for the citizens and the public officers to improve personal safety and prevent crimes. These services will include:

- the automatic monitoring of pedestrian subways by means of mobile and low-power audio and proximity sensors;
- the automatic monitoring of traffic scenes by cameras for data collection and web-based delivery of traffic news to citizens;
- the generation of a feedback in pedestrian crossing systems to select the best duration of the green signal for the crossing;
- the automatic monitoring of public parks with a plethora of cameras (both fixed and PTZ).

In this paper, we focus on this last scenario. In this framework, ubiquitous computer vision systems can be exploited to give both services to citizens and support to public officers in crime prevention and forensics.

The second issue is straightforward and it is the primary goal of (automatic or not) surveillance systems. The possible applications of automatic monitoring systems are manifold: for example, the detection of a person that repeatedly moves along a path close to a landmark (near a children playground or a security-enforced place) could be annotated for offline further investigation. The detection of a dangerous behaviour of people near children, or people leaving an object (for example, underneath a bench) could be important for a fast reaction and alerting system.

At the other hand, citizen could be interested to multimedia and web services to access in real time and "visit" public places like the parks, take a look at the weather condition or the presence of crowd. However, all the above-mentioned services, and more than ever the public park one, must deal with privacy issues. The laws in Italy are very restrictive and do not allow the transmission of the person's identity (e.g. the face) without an explicit permission. Thus, another key aspect of our research is in finding automatic techniques for extracting faces from the image and obscuring them in real time.

For detecting all of these situations, a multi-camera setup must be considered. In fact, the availability of multiple points of view and the redundancy of information allow a more precise (though not unfailing) monitoring of complex scenes.

In particular, two problems can be (partially) solved by multiple cameras: the coverage of wide areas (by means of either overlapped or non-overlapped cameras) and the management of object occlusions (by exploiting the different viewpoints). The merge of the data provided by multiple cameras poses, however, some problems too. A main problem is that the identity of the objects moving from one camera to another must be preserved, in order to analyze their behaviours over the whole scene. This process is known in the literature as *"consistent labelling"* and becomes challenging when cameras can not be manually calibrated.

To test our algorithms we created a test bed on our campus, installing four partially overlapped cameras (three fixed and one PTZ) as sketched in Fig. 1, in a zone where many people are passing through, there are some benches and the illumination conditions are typical of an outdoor environment.

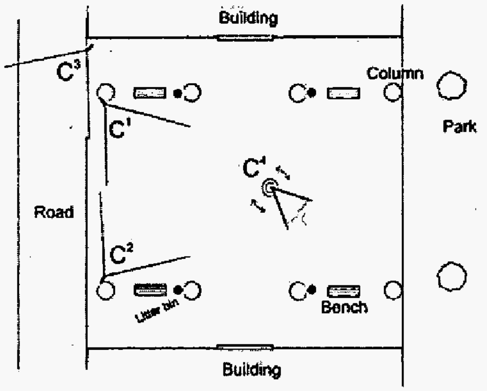


Fig. 1. Sketch of the test bed.

In the rest of the paper, we report on a novel approach for consistent labelling with automatic learning of calibration parameters, so as to work in a mosaic image based on homography. Then, we describe the techniques adopted to track multiple people and provide a semantic adaptation of the video with face obscuration.

2. CONSISTENT LABELLING

The problem of consistent labelling has been addressed in the literature mainly in two ways. The first relies only on the appearance of the objects, the second on the geometrical relationship between overlapped cameras.

The appearance-based approaches base the matching essentially on the colour of the tracks, by using invariants to illumination changes and texture features, and clustering based on mean shift (Li et al. (1)) or matching of colour histograms (Krumm et al. (2)). However, 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 have been based on geometrical constraints. Geometry-based approaches can be further subdivided into calibrated (Mittal and Davis (3), Yue et al. (4)) and uncalibrated (Khan and Shah (5)) approaches. The approach in (5) is based on the computation of the so-called *Edges of Field of View*, i.e. the lines delimiting the field of view of each camera and, thus, defining the overlapped regions. There are also approaches that try to mix information about the geometry and the calibration with those provided by the visual appearance, such as Kang et al. (6), or Chang and Gong (7).

The proposed solution relies on the creation of the so-called *edges of field of view* (EOFOV, hereinafter) to automatically calibrate cameras. An EOFOV line is the projection on the camera C^j of the line s of the camera C^i and it is denoted by $L^{i,s}_j$. Each line $L^{i,s}_j$ divides the image on camera C^j into two half-planes, one overlapped with camera C^i and the other disjoint. The intersection of the overlapped semi-planes defined by the EOFOV lines from camera C^i to camera C^j defines the overlapping area Z^i_j.

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^i_k the object segmented and tracked with label k in the camera C^i and SP^i_k the point of contact with the ground plane (*support point*, hereinafter). The support point can be easily 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 in camera C^j and tracked with label p, it is directly associated to O^i_k. Therefore, in this moment (known as the moment of the "camera handoff"), the support point SP^i_k can be associated to SP^j_p (if it is visible). In this case the point SP^i_k lies on the EOFOV line $L^{j,s}_i$ for the camera C^i. The equation of each line $L^{j,s}_i$ is computed by collecting a set of coordinates of the support point SP^i_k detected at the camera handoff and exploiting a Least Square optimization (Fig. 2.a).

In the method proposed in (5), the support points are extracted at the camera handoff moment. This can bring to false correspondences, as in the case of a person entering from the bottom of the image (Fig. 2.b). In such a situation the support point can not be simply computed by taking the lowest point of the detected blob because the requirement to have both the support points lying on the ground plane is not verified.

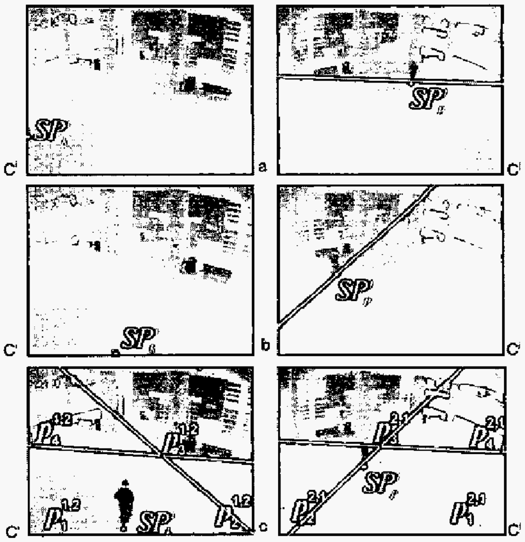


Fig. 2. Examples of EOFOV computation.

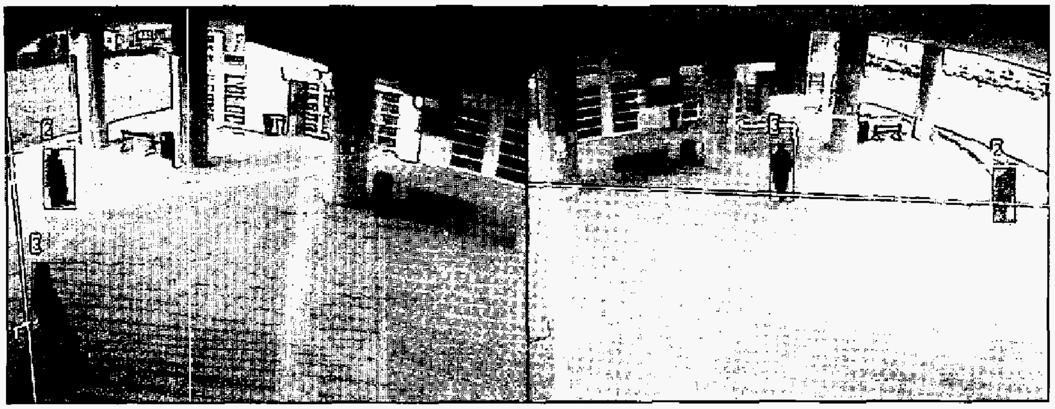


Fig. 3. Examples of simultaneous transition.

To solve this problem, we modified the approach of Khan and Shah by delaying the computation of the EOFOV lines to the moment in which the object is completely entered the scene of the new camera (see Fig. 2.c). This can bring to a displacement of the line with respect to the actual limit of the image, but it assures the correct match of the feet's position in the two views. As a consequence, the actual FOV lines (displayed in green in Fig. 2.c) 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 in (5) establishes the consistent labelling only in the exact moment of the camera handoff from C^i to C^j.

This approach has two main limits: if two or more objects cross simultaneously (Fig. 3) an incorrect labelling can be established; if they are merged from the view of C^j at the camera handoff, but then they separate, the consistent labelling with the labels of C^i can not be recovered (Fig. 4).

We propose to overcome these problems by means of homography, thus extending the matching search to the whole zone of overlap of 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 $P_j^i = \left\{ p_1^{i,j}, p_2^{i,j}, p_3^{i,j}, p_4^{i,j} \right\}$ and $P_i^j = \left\{ p_1^{j,i}, p_2^{j,i}, p_3^{j,i}, p_4^{j,i} \right\}$, where the subscripts indicate corresponding points in the two cameras (see Fig. 2.c). These four associations between points of the camera C^i and points of the 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 be easily obtained with the equation $H_i^j = \left(H_j^i \right)^{-1}$.

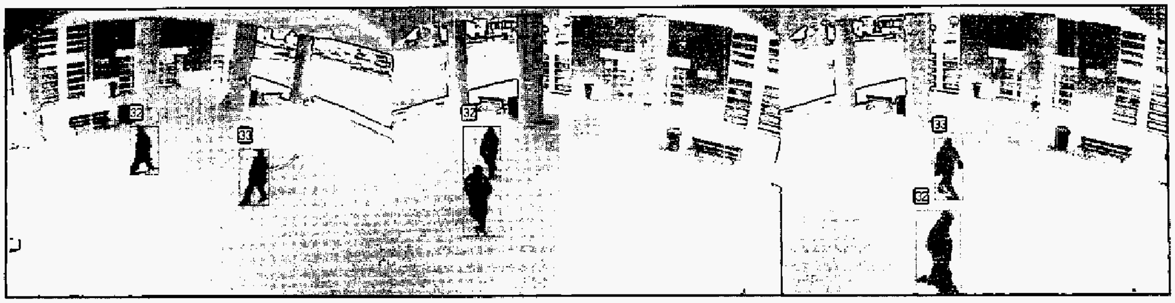
Each time a new object O_k^i is detected in the camera C^i in the overlapping area (not only at the moment of the camera handoff), its support point SP_k^i is projected in C^j by means of the homographic transformation. Called $\left(x_{SP_k^i}, y_{SP_k^i} \right)$ the coordinates of the support point SP_k^i, we can write the projected point in homogeneous coordinates $\left[a,b,c \right]^T = H_j^i \left[x_{SP_k^i}, y_{SP_k^i}, 1 \right]$. The projected point $\widetilde{SP_k^j}$ corresponds on the image plane of C^j to the projective coordinates $\widetilde{x^j} = {}^a\!/_c$ and $\widetilde{y^j} = {}^b\!/_c$. 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 in the 2D plane from these coordinates:

$$O_k^i \longleftrightarrow O_p^j \left| p = \arg\min_q D\left(\widetilde{SP_k^j}, SP_q^j \right) \right. \quad \forall q \in \mathbf{O}^\mathbf{j} \qquad (1)$$

where $D(\cdot)$ denotes the Euclidean distance and $\mathbf{O}^\mathbf{j}$ is the set of objects detected in C^j. The results achieved with this approach in the two cases above reported are shown in Fig. 3 and Fig. 4.c, respectively, where the correct label assignment is achieved.

3. FACE DETECTION AND TRACKING

Face detection is a widely explored research area in computer vision. Two recent surveys, Yang et al. (8) and Hjelm and Low (9), collect a large number of proposals about face detection. Most of them are based on a skin colour detection (Jones and Rehg (10)) followed by a face candidate validation achieved exploiting geometrical and topological constraints.



(a) C^2 frame 1250 (b) C^1 frame 1250 (c) C^1 frame 1260

Fig. 4. Examples of merged transition.

Unfortunately, most of the colour-based approaches are very expensive from the computational point of view 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 as the one proposed in Birchfield (11) can be adopted. A different approach, instead, is the one proposed by Maio and Maltoni in (12), which works on grey scale images. In particular, the face candidates are obtained through ellipse detection applied over the gradient.

The algorithm we adopt is based on the elliptical approximation of the head shape and the generalized Hough transform on tracked people. For tracking we use an appearance based approach as described in Cucchiara et al. (13). When people move in zones of overlapped field of view of two or more cameras, the previously described algorithm checks the consistency. The developed method exploits and improves the best ideas proposed in (11) and (12). The first uses both colour and gradient information but the search of the head is limited to a neighbourhood of a predicted position. The problem of this solution is that it needs a frame rate too high to make reliable predictions. Instead, (12) adopts a solution based on the elliptical Hough transform; differently from the previous, this solution does not require any tracking nor prediction, because the processing of each frame is stand-alone. As in (13), a face colour histogram must be available as a model. To this aim, a supervised learning phase is performed to compute a histogram of skin and hair colours. Thus, for each tracked object, two different Hough transforms are computed: one gradient-based and one colour-based. The points belonging to the edges of the track (obtained with Canny edge detectors) vote for the first transform according with the gradient value. The selection of the voted pixels is done by moving on the image in the same direction of the gradient with a distance obtained from the estimated head size.

Similarly, a point of the object votes for the colour-based transform if its colour has a non-zero value on the histogram of the saved head model.

In this case it votes for all the points inside an ellipse having the same size of the head and the actual pixel as the centre, and the rate is proportional to the model histogram value corresponding to the colour of the pixel. After that, 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 head centre. Finally, the head is obscured, as in Fig. 5. The proposed algorithm requires a minimum size of the head, but, as it is possible to see in Fig. 6, if the head is too small, the privacy is not a problem, since the person can not be recognized anyway.

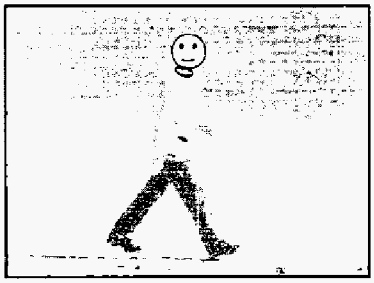


Fig. 5. Example of face obscuration.

Fig. 6. Case of avoidable face obscuration.

4. EXPERIMENTS AND CONCLUSIONS

Even if the project is in its starting phase only, we can give reliable experimental results of our techniques, thanks to the set up of camera and vision system available in our campus. People tracking in a single camera is robust enough and is not sensitive to track overlaps, posture changes, track occlusions due to static objects (e.g., columns) (see (13)).

The new consistent labelling algorithm has been tested extensively with two partially overlapped cameras. A mosaic image can be obtained blending an undistorted image from a camera with the homographical projection of the other one, and the complete trajectory of people can be drawn (Fig. 7). This trajectory can be further analyzed to detect, for example, suspicious situations.

Finally, an automatic video adaptation can be carried out by modifying the video content with an on-line face obscuration to meet privacy constraints.

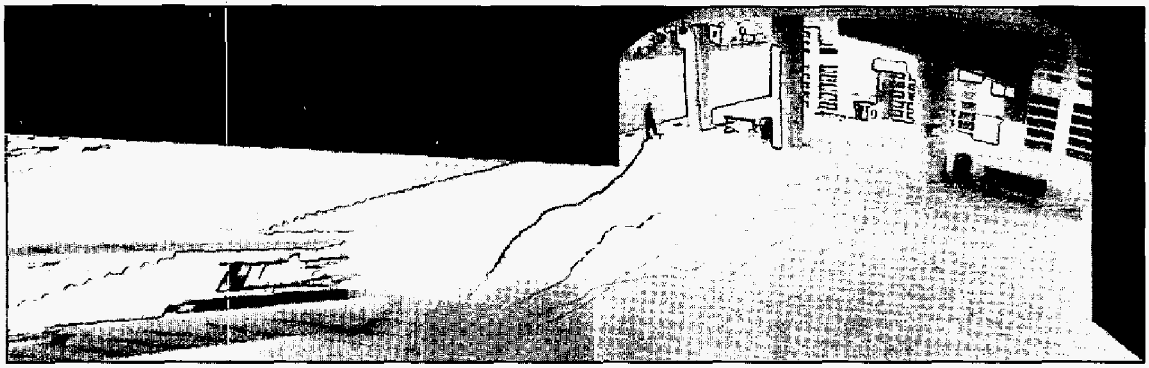


Fig. 7. Example of mosaic image with trajectory.

REFERENCES

1. Li J., Chua, C.S., Ho, Y.K. 2002. "Color based multiple people tracking", Proc. of IEEE Intl Conf on Control, Automation, Robotics and Vision, 1, 309-314

2. Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., Shafer, S. 2000. "Multi-camera multi-person tracking for EasyLiving", Proc. of IEEE Intl Workshop on Visual Surveillance, 3-10

3. Mittal, A., Davis, L. 2001. "Unified multi-camera detection and tracking using region-matching", Proc of IEEE Intl Workshop on Multi-Object Tracking, 3-10

4. Yue, Z., Zhou, S.K., Chellappa, R., 2004. "Robust two-camera tracking using homography", Proc. of IEEE Intl Conf on Acoustic, Speech and Signal Processing, 3, 1-4

5. Khan, S., Shah, M., 2003. "Consistent labelling of tracked objects in multiple cameras with overlapping fields of view", IEEE Trans on Patt Anal Mach Int, 25(10), 1355-1360

6. Kang, J., Cohen, I., Medioni, G., 2003. "Continuous tracking within and across camera streams", Proc. of IEEE Intl Conf on Computer Vision and Pattern Recognition, 1, 267-272

7. Chang, S., Gong, T.-H., 2001. "Tracking multiple people with a multi-camera system", Proc. of IEEE Workshop on Multi-Object Tracking, 19-26

8. Yang, M., Kriegman, D.J., Ahuja, N., 2002. "Detecting Faces in Images: A Survey" IEEE Trans on Patt Anal Mach Int, 24(1), 34-58.

9. Hjelm, E., Low, B.K., 2001. "Face Detection: A Survey," Computer Vision and Image Understanding, 83(3), 236-274.

10. Jones, M.J., Rehg, J.M., 2002. "Statistical Color Models with Application to Skin Detection," Intl Journal of Computer Vision, 46(1), 81-96.

11. Birchfield, S., 1998. "Elliptical Head Tracking Using Intensity Gradients and Color Histograms," Proc. of IEEE Intl Conf on Computer Vision and Pattern Recognition, 232-237.

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

13. Cucchiara, R., Grana, C., Tardini, G., Vezzani, R., 2004. "Probabilistic People Tracking for Occlusion Handling" Proc of Intl Conf on Pattern Recognition, 1, 132-135.