Article

A multi-camera vision system for fall detection and alarm generation

Rita Cucchiara,[1] Andrea Prati[2] and Roberto Vezzani[1]

(1) Dipartimento di Ingegneria dell'Informazione, University of Modena and Reggio Emilia, Via Vignolese 905/b, 41100 Modena, Italy

(2) Dipartimento di Scienze e Metodi dell'Ingegneria, University of Modena and Reggio Emilia, Via Amendola 2, 42100 Reggio Emilia, Italy

E-mail: cucchiara.rita@unimore.it

Abstract: *In-house video surveillance can represent an excellent support for people with some difficulties (e.g. elderly or disabled people) living alone and with a limited autonomy. New hardware technologies and in particular digital cameras are now affordable and they have recently gained credit as tools for (semi-) automatically assuring people's safety. In this paper a multi-camera vision system for detecting and tracking people and recognizing dangerous behaviours and events such as a fall is presented. In such a situation a suitable alarm can be sent, e.g. by means of an SMS. A novel technique of warping people's silhouette is proposed to exchange visual information between partially overlapped cameras whenever a camera handover occurs. Finally, a multi-client and multi-threaded transcoding video server delivers live video streams to operators/remote users in order to check the validity of a received alarm. Semantic and event-based transcoding algorithms are used to optimize the bandwidth usage. A two-room setup has been created in our laboratory to test the performance of the overall system and some of the results obtained are reported.*

Keywords: multi-camera, posture classification, semantic transcoding

1. Introduction

The importance of health care, especially for elderly and disabled people, is obvious and fully demonstrated. For example, in the last eight years we have seen the huge interest of the European Commission in research on the topic of 'Ambient assisted living (AAL) for the ageing society'. As a result of this interest, Call 6 of the Sixth Framework Programme[1] (just finished) and the next Seventh Framework Programme of the Commission will fund projects on the topic. Developing systems and tools for easing the transition phase from dependent to independent living in an acceptable way should be the main goal of these projects.

The increase each year in deaths and injuries in domestic incidents has shown up in-house safety as an emerging field of research. This is even more true in the case of the elderly and disabled, for whom normal everyday tasks can be difficult and a continuous source of injury and danger. Existing technological solutions basically aim at either preventing injuries or communicating (to the health care provider or to a relative) problems by means of worn sensors. For instance, fall detection can be achieved with electronic devices to be worn or kept in a pocket by the user. Since these systems are typically based on low-level sensors, they are

[1] *http://cordis.europa.eu/ist/so/aal/home.html.*

somewhat inaccurate in sensing the environment, they can be too intrusive and they can easily be forgotten by the elderly/disabled user.

To overcome these drawbacks, cameras are increasingly included in in-house safety systems. In fact, cameras are sensors that can be used for detecting different events simultaneously and they are less intrusive since they do not require the wearing of any sensors. Moreover, in general, the data provided by cameras are semantically richer and more accurate than from standard sensors. Unfortunately, data processing requires advanced computer vision techniques that are prone to errors and computationally expensive. Moreover, cameras are bounded in their field of view, i.e. their accuracy is limited by occlusions and directly correlated to the accuracy of the acquisition CCD sensors. In addition, in the case of house monitoring, it is not possible to use a single camera to monitor all the rooms; multiple cameras need to be used and coordinated.

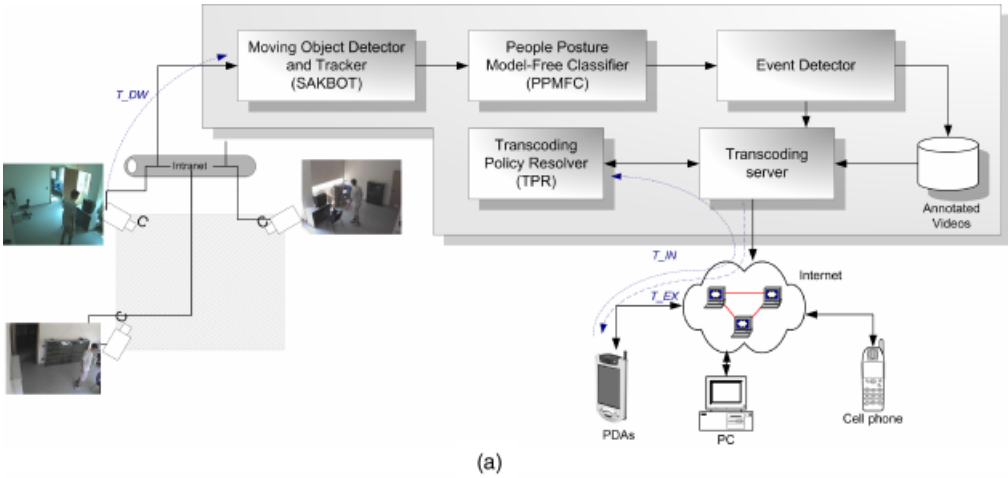
Despite these difficulties, cameras and computer vision have recently gained credit as tools for (semi-)automatically assuring people's safety. Among the different activities to be monitored for non-completely-independent people living alone at home, falls are of particular importance for obvious reasons. Using computer vision, falls can be detected by analysing the person's posture and detecting sudden changes in posture (e.g. from standing to lying).

Another important issue to be considered is the best action to be taken when a fall (or, in general, an interesting event) is detected. Using cameras, live video streams can be delivered to the operator/remote user to help him in understanding what happened. In addition, ubiquitous access by means of mobile devices (such as a PDA or last-generation cell phone) to assess the situation is unquestionably desirable. To make this possible, SMS is used to issue notification of the alarm to the device and video streaming must be adapted to the limited resources (in terms of displaying capabilities, computational power and available bandwidth for the network connection) of the mobile device.

This paper presents our approach to address all the issues reported above. The overall scheme of the system is shown in Figure 1. Our approach exploits computer vision techniques to detect and track people inside a single room with a single camera and proposes a new method based on hidden Markov models (HMMs) for classifying posture and detecting falls. Additionally, multiple cameras are used to cover different rooms and the camera handoff is treated by warping the person's appearance in the new view by means of homography. Finally, alarms are generated when a person falls and an SMS message is sent to a PDA with the link for live video streaming. The paper is structured as follows. The next section will report the main contributions in the field of posture classification by means of computer vision and automatic systems for elderly health care. Sections 3 and 4 describe in detail the single camera and multi-camera modules, respectively, while Section 5 describes the approach used to communicate alarms to the operators. The experiments and conclusions are reported in Section 6.

2. Related work

Many methods based on computer vision have been proposed in the literature to classify people's posture. They can be differentiated by the more or less extensive use of a two-dimensional or three-dimensional model of the human body (Moeslund & Granum, 2001). In accordance with this, most of them can be classified into two basic approaches to the problem. From one side, some systems (such as that proposed by Haritaoglu *et al.*, 1998) use a direct approach and base the analysis on a detailed human body model. In many of these cases, an incremental predict–update method is used, retrieving information from every body part. These approaches are often too constrained to the human body model, resulting in unreliable behaviour in the case of occlusions and perspective distortions that are very common in cluttered, relatively small, environments like a room. Moreover, for in-house surveillance systems, low-cost



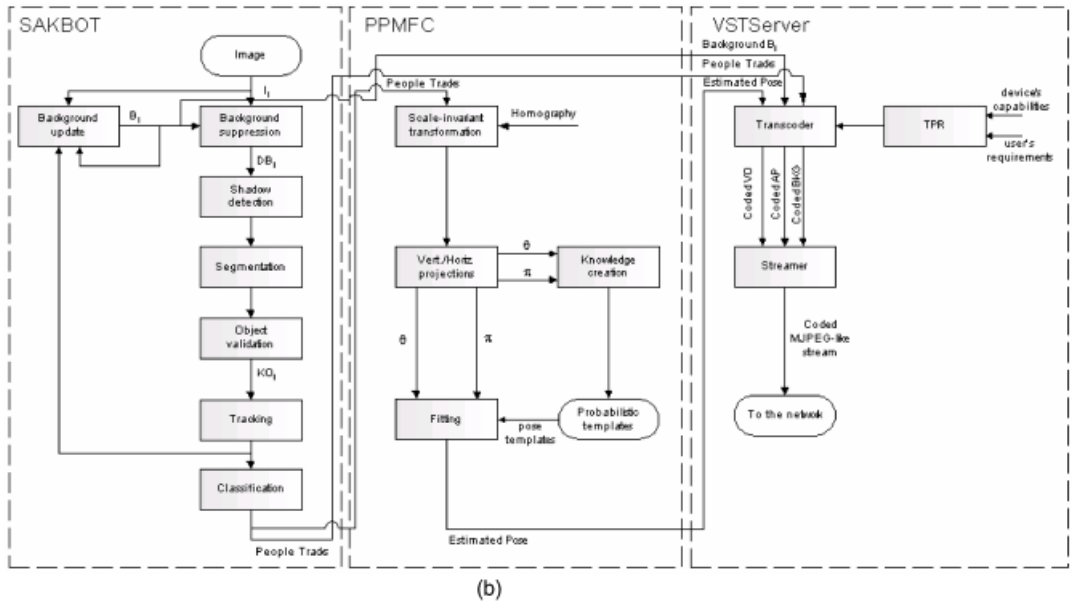


Figure 1: *(a) Overall scheme of the system and (b) details of the three main modules.*

solutions are preferable; thus stereovision or three-dimensional multi-camera systems should be discarded. Consequently, algorithms of people's posture classification should be designed to work with simple visual features from a single view, such as silhouettes, edges and so on (Haritaoglu *et al.*, 1998; Moeslund & Granum, 2001).

Since posture is often related to the silhouette, occlusions are very critical. Even though some probabilistic approaches are able to maintain the shape in the case of occlusions (as in Haritaoglu *et al.*, 1998) they are likely to fail if the occlusion occurs at the beginning, i.e. when the track is first created. This problem can be solved by the use of multiple cameras. Moreover, the need for multiple points of view and a distributed system is required in order to cover the entire environment (e.g. the house) and continuously track people in it.

Unfortunately, the possibility of having full cover of the environment and solving occlusions

does not come free. The technical problems in multiple camera systems are several and they have been summarized in Hu *et al.* (2004) into six classes: installation, calibration, object matching, switching, data fusion, and occlusion handling. Among these, object matching is the problem most addressed in the literature and it provides the basic tools also for occlusion handling.

Several proposals have been made to maintain the correspondence of the same tracked object during a camera handoff. Most of these require a partially overlapping field of view (Cai & Aggarwal, 1999); others use a feature-based probabilistic framework to maintain a coherent labelling of the objects (Khan & Shah, 2003). All these papers aim at keeping correspondences between the same tracked object, and none of them is capable of handling occlusions during the camera handoff phase.

Approaches to multi-camera tracking can be generally classified into three categories: geometry-based, colour-based, and hybrid approaches. The first class can be further subdivided into calibrated and uncalibrated approaches. A particularly interesting paper on the calibrated approach is by Yue *et al.* (2004) in which homography is exploited to solve occlusions. Single camera processing is based on particle filters and on probabilistic tracking based on appearance to detect occlusions. Once an occlusion is detected, homography is used to estimate the track position in the occluded view, by using the track's last valid positions and the current position of the track in the other view (properly warped in the occluded one by means of the transformation matrix). A relevant example of the uncalibrated approach is the work of Khan and Shah (2003). Their approach is based on 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. Through a learning procedure in which a single track moves from one view to another, an automatic procedure computes these edges that are then exploited to keep consistent labels on the objects when they pass from one camera to the adjacent camera.

Colour-based approaches base the matching essentially on the colour of the tracks, as in Li *et al.* (2002) where a colour space invariant to illumination changes is proposed and histogram-based information at low (texture) and mid (regions and blobs) level are exploited to solve occlusions and match tracks by means of a modified version of the mean shift algorithm.

Hybrid approaches mix information about the geometry and the calibration with those provided by the visual appearance. These methods use probabilistic information fusion (Kang *et al.*, 2003) or Bayesian belief networks (Cai & Aggarwal, 1999).

Our approach is similar to that proposed by Yue *et al.* (2004) but, differently from them, appearance models of the tracks are warped from one view to another using not the ground plane but a vertical plane passing from the person's feet and triggered by an external or internal input. We will also report results of an experiment that aims at analysing the limits of the approach depending on the amount and type of the occlusion that has occurred.

3. Single camera module

In our multi-camera system, moving objects are extracted from each camera by exploiting background suppression with selective and adaptive update in order to quickly react to changes and also to take 'ghosts' (i.e. aura left behind by an object that begins to move) into account, using the SAKBOT system shown in Figure 1 and described by Cucchiara *et al.* (2003b). After the object extraction, a sophisticated tracking algorithm is used to cope with occlusions and split/merge of objects. A probabilistic and appearance-based tracking, similar to that proposed by Senior (2002), is used to handle objects with non-rigid motion, variable shapes (like people) and frequent occlusions. This tracking algorithm maintains, in addition to the current blob, the appearance image (or temporal template) and the probability mask of the track. Appearance image is obtained with a temporal integration of the colour images of the blobs, while the probability mask associates with each

point of the map a probability value that indicates its reliability. Comparing the current blob with the appearance image of the tracks it is also possible to detect if the person is subject to an occlusion or not (Cucchiara *et al.*, 2005).

Finally, tracks that satisfy some geometrical and colour constraints are classified as people and submitted to the posture classifier. Four main postures are considered: standing, crawling, sitting and lying. For this, we exploit a classifier based on the projection histograms computed over the blobs of the segmented people. The projection histograms $\mathrm{PH} = (\vartheta(x); \pi(y))$ describe the way in which the silhouette's shape is projected on the x and y axes. Since the projection histograms depend on the blob size, and consequently on the position of the person inside the room, we first scale them according to the distance of the person with respect to the camera. To compute this distance, feet detection and tracking module together with a homography relation obtained through camera calibration are exploited (Cucchiara *et al.*, 2005).

Although projection histograms are very simple features, they have proven to be sufficiently detailed to discriminate between the postures we are interested in. However, this classifier is precise enough if the lower level segmentation module produces correct silhouettes. By exploiting knowledge embedded in the tracking phase, many possible classification errors due to the imprecision of the blob extraction can be corrected. In particular, to deal with occlusions and segmentation errors due to noise, the projection histograms are computed over the temporal probabilistic maps obtained by the tracker instead of the blobs extracted frame by frame. Despite the improvements given by the use of appearance mask instead of blobs, a frame-by-frame classification is not reliable enough. However, the temporal coherence of the posture can be exploited to improve the performance: in fact, the person's posture changes slowly and through a transition phase during which its similarity with the stored templates decreases. With this aim, an HMM formulation has been adopted. Using the notation proposed by Rabiner (1989) in his tutorial, the following sets are defined:

- the state set S, composed of N states: $S = \{S_1, \ldots, S_N\} = Main\_Postures$;
- the initial state probabilities $\Pi = \{\pi_i\}$ set equal for each state ($\pi_i = 1/N$) – the choice of the values assigned to the vector Π affects the classification of the first frames only, and then it is not relevant;
- the matrix $\boldsymbol{A}$ of the state transition probabilities, computed as a function of a reactivity parameter α (empirically determined; e.g. we set $\alpha = 0.95$ during our experiments). The probabilities to remain in the same state and to pass to another state are considered equal for each posture. Then, the matrix $\boldsymbol{A}$ has the following structure:

$$\boldsymbol{A} = \boldsymbol{A}(\alpha) = \{A_{ij}\}$$
$$A_{ij} = \begin{cases} \alpha & i = j \\ \dfrac{1-\alpha}{N-1} & i \neq j \end{cases} \quad (1)$$

The observation symbols and the observation symbol probability distribution $\boldsymbol{B}$ have to be defined. To this aim we can use the set of possible projection histograms as observation symbols, since it is numerable – but that means the computation of a very large matrix, composed of N rows and w^h columns (where w and h are the sizes of the images). Thus, we prefer to compute directly the probability values b_j, which indicate the probability of having a particular observation (histogram) belonging to the state (posture) j, through the output of the frame-by-frame classifier:

$$b_j = P_j = P(\widehat{\mathrm{PH}} \,|\, \mathrm{posture} = S_j) \quad (2)$$

The HMM presented does not require any additional training phase because it exploits the probability maps directly. Then, at each frame, the probability of being in each state is computed with the traditional forward algorithm. Finally, the HMM input has been modified to keep account of the visibility status of the person. In fact, if the person is completely

occluded, the reliability of the posture must decrease with time. In such a situation, it is preferable to set $b_j = 1/N$ as the input of the HMM. Hence the state probabilities tend to a uniform distribution (that models the increasing uncertainty) with a delay that depends on the previous states: the higher the probability of being in a state S_j, the higher the time required to lose this certainty. To manage the two situations simultaneously and to cope with intermediate cases (i.e. partial occlusions), a generalized formulation of the HMM input is defined:

$$b_j = P(\widehat{\mathrm{PH}}|S_j)\frac{1}{1+n_{\mathrm{fo}}} + \frac{1}{N}\frac{n_{\mathrm{fo}}}{1+n_{\mathrm{fo}}} \qquad (3)$$

where n_{fo} is the number of frames for which the person is occluded. If n_{fo} is zero (i.e. the person is visible), b_j is computed as in equation (2); otherwise, the higher the value of n_{fo}, the more it tends to a uniform distribution.

In Figure 2, the benefits of the introduction of the HMM framework are clear. The results are related to a video in which a person passes behind a stack of boxes always in a standing position. During the occlusion (highlighted by the grey strips) the frame-by-frame classifier fails (it states that the person is lying). Instead, through the HMM framework, the temporal coherence of the posture is preserved, even if the classification reliability decreases during the occlusion.

4. Multi-camera module

As stated above, the probabilistic tracking is able to handle occlusions and segmentation errors in the single camera module. However, the strong hypothesis to be robust to occlusions is that the track has been seen for some frames without occlusions in order for the appearance model to be correctly initialized. This hypothesis is erroneous when the track is occluded since its creation (as in Figure 3(b)).

Our proposal is to exploit the appearance information from another camera (where the track is not occluded) to solve this problem. If a

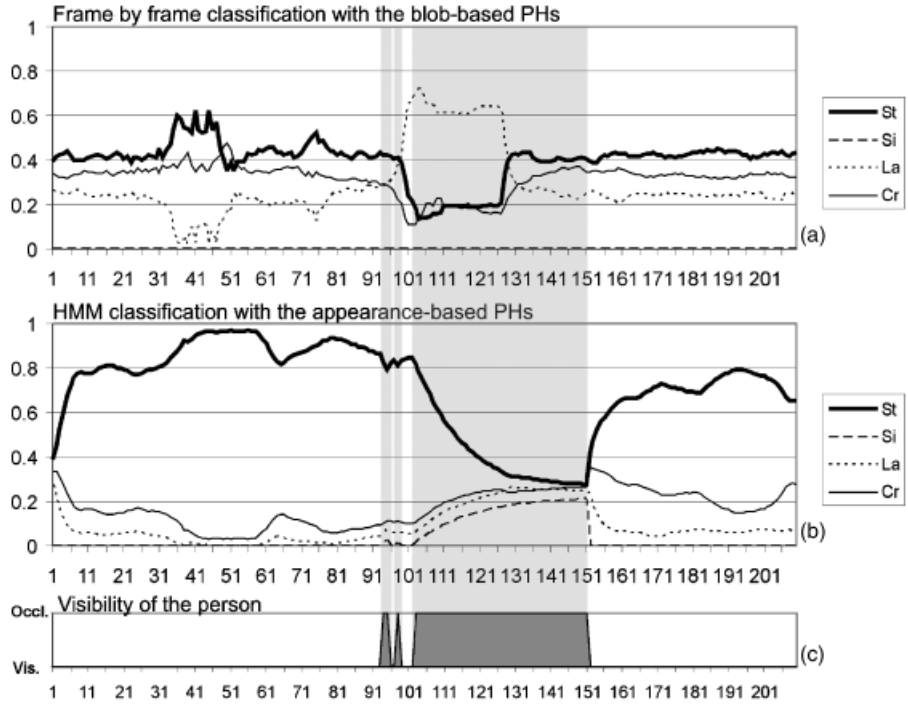


Figure 2: *Frame-by-frame and HMM posture classification during a strong occlusion.*

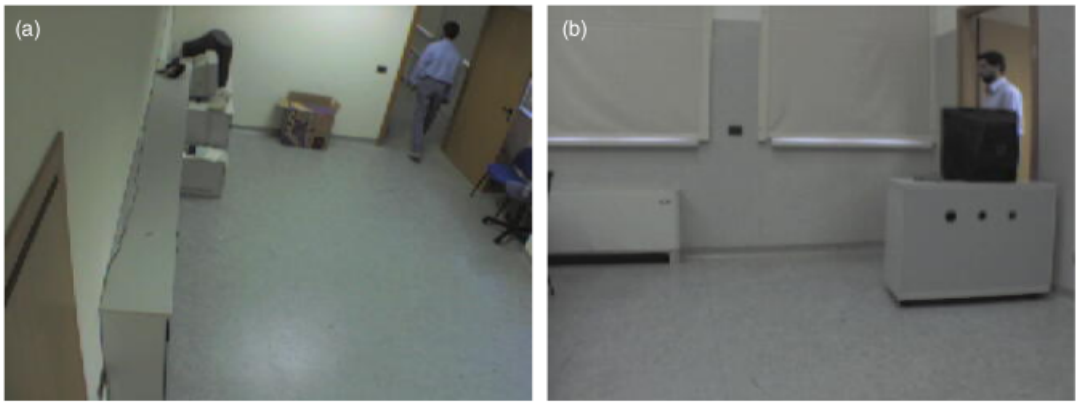


Figure 3: *An example of track occlusion during its creation: (a) the source and (b) the destination view.*

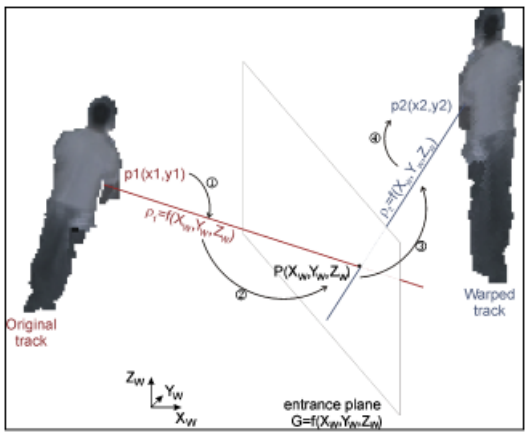


Figure 4: *Track warping: (1) exploiting the calibration of camera 1; (2) intersection with the entrance plane; (3) calibration of camera 2; (4) intersection with the camera 2 plane.*

person passes between two monitored rooms, it is possible to keep the temporal information stored into the track extracted from the first room (Figure 3(a)) and use it to initialize the corresponding track in the second room (Figure 3(b)).

With this aim, with reference to Figure 4, the following assumptions are used.

- The two cameras are calibrated with respect to the same coordinate system (X_W, Y_W, Z_W).
- The equation of the plane $G = f(X_W, Y_W, Z_W)$ containing the entrance is given.
- It is possible to obtain the exact instant t_{SYNC} when the person passes into the second room. To do this, the three-dimensional position of the support point SP of the people inside the room could be used, or otherwise a physical external sensor could be adopted for a more reliable trigger.
- All the track points lie on a plane P parallel to the entrance plane G and containing the support point SP (hereinafter we call P the person's plane). This assumption holds if the person passes the entrance in a posture such that the variance of its points with respect to the direction normal to P is low enough (e.g. standing posture).

The first three assumptions imply only an accurate installation and calibration of cameras and sensors, while the last one is a necessary simplification to warp the track between the two points of view. Under this condition, in fact, the three-dimensional position of each point belonging to the appearance image of the track can be computed and then its projection on a different image plane is obtained (Figure 4).

In particular, the process mentioned above is applied only to the four corners of the tracks and thus the homography matrix $\boldsymbol{H}$ that transforms each point between the two views can be computed:

$$[\,x_2 \quad y_2 \quad 1\,]^{\mathrm{T}} = \boldsymbol{H}_{3\times 3} \cdot [\,x_1 \quad y_1 \quad 1\,]^{\mathrm{T}} \qquad (4)$$

Through $\boldsymbol{H}$ it is possible to re-project both the colour components and the α value of each track point from the point of view of leaving the room to the point of view of entering the room (Figure 3). The re-projected track is used as initialization for the new view that can in such a manner solve the occlusion by continuing to detect the correct posture.

5. Alarm generation

As depicted in Figure 1, the system is built on the client–server architecture. With this aim, we implemented a multi-client and multi-threaded *transcoding video server* called VSTServer (video streaming transcoding server). Among the different threads present in the system, three are critical. A downloading thread (T_{DW}) is devoted to acquiring the sequence of images from the network camera in the streaming mode. An inquiring thread (T_{IN}) establishes the communication between client and server and sets the transcoding policies. Whenever the initial parameters (requests of size, bandwidth etc.) are set, the connection between client and server is passed to an execution thread (T_{EX}). From this moment, another client can connect to the server. The threads are decoupled to allow the maximum frame rate in getting the image from the camera, despite the possible slowdowns due to slow clients. The communication between the two threads is based on shared buffers (in which the T_{DW} puts the image and from which the T_{EX} picks it up), with a semaphore-based protocol to obtain synchronization between the two threads.

As mentioned above, images coming from the cameras are processed to detect dangerous situations or events. The corresponding alarms (in our case the person's falls) can be managed in several ways. For example, a control centre can be advised and connected through a video–audio link with the assisted person. Obviously, all the events can be saved on a database for further processing. In addition, a vocal message or an SMS can be sent to a relative or a neighbour on their cell phone or PDA, and in this case a link for a low-bandwidth video connection to assert a person's condition can be provided. In this context, due to the limited computational, storage and display capabilities of the mobile device and to the probable low-bandwidth connection, a *video adaptation* is mandatory. Typically, a syntactic adaptation of the video (by means of frame size reduction, frame skipping or quality deterioration) is used. Our claim is that in extreme applications like this, in which at one side the connection bandwidth can be very limited (e.g. in the case of the General Packet Radio Service) and, at the other side, a good image quality can save lives, normal only-syntactic downloading or transcoding methods are not effective. In previous work (Cucchiara *et al.*, 2003a), we proposed *semantic-based*

Figure 5: *Example of transcoded video: left, original uncompressed frame; right, semantic-based video adaptation in which the fallen person is sent with higher quality to allow identification.*

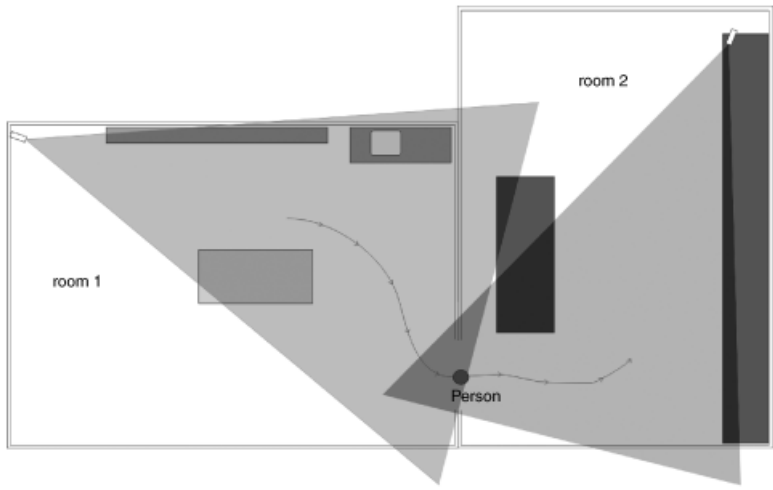


Figure 6: *A scheme of the two rooms used for our tests.*

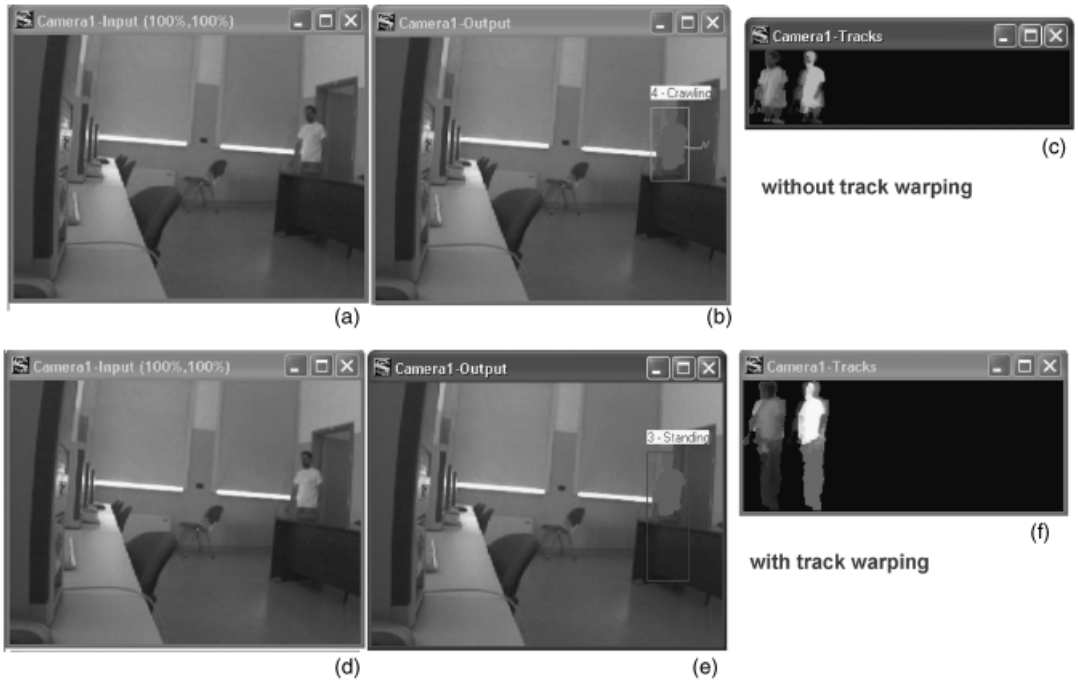


Figure 7: *Initial occlusion resolved with track warping: (a), (d) input frames; (b), (e) output of the posture classification; (c), (f) appearance images (left) and probability maps (right).*

techniques for video content adaptation. The rationale is that if we know (from the user or automatically) which are the relevant semantics in the video context, we can exploit this to selectively transcode the video: the bandwidth saved by degrading the irrelevant contents can be used to increase the quality of the relevant contents. The user can associate different weights depending on the relevance of each part of the video: the higher the weight, the more relevant must be considered the class, and the transcoding policy resolver (see Figure 1)

will apply a less aggressive set of transcoding policies.

At the client side, the stream is decoded and the resulting video is visualized on the mobile device by applying further scaling and general adaptation to the display capabilities. An example of different compression applied to different areas of the image resulting in a transcoded video is shown in Figure 5. Further details can be found in Cucchiara *et al*. (2003c).

6. Experiments and conclusions

As a test bed for our multi-camera system, a two-room setup has been created (Figure 6). The two rooms share a door equipped by an optical sensor used to trigger the passage of people. We have taken several videos of transition between the first and the second room. Furthermore, in the second one we have placed various objects between the door and the camera to simulate different types and amounts of occlusions. In particular, occlusions starting from both the bottom part and the middle part of the body have been created. In the case of the videos of the first type (bottom occlusions), the single camera posture classifier tends to fail because the body shape is incomplete and the feet are not visible to be tracked. An example of erroneous classification if the track warping is disabled is reported in the upper row of Figure 7. The lower row, instead, reports the result achieved by

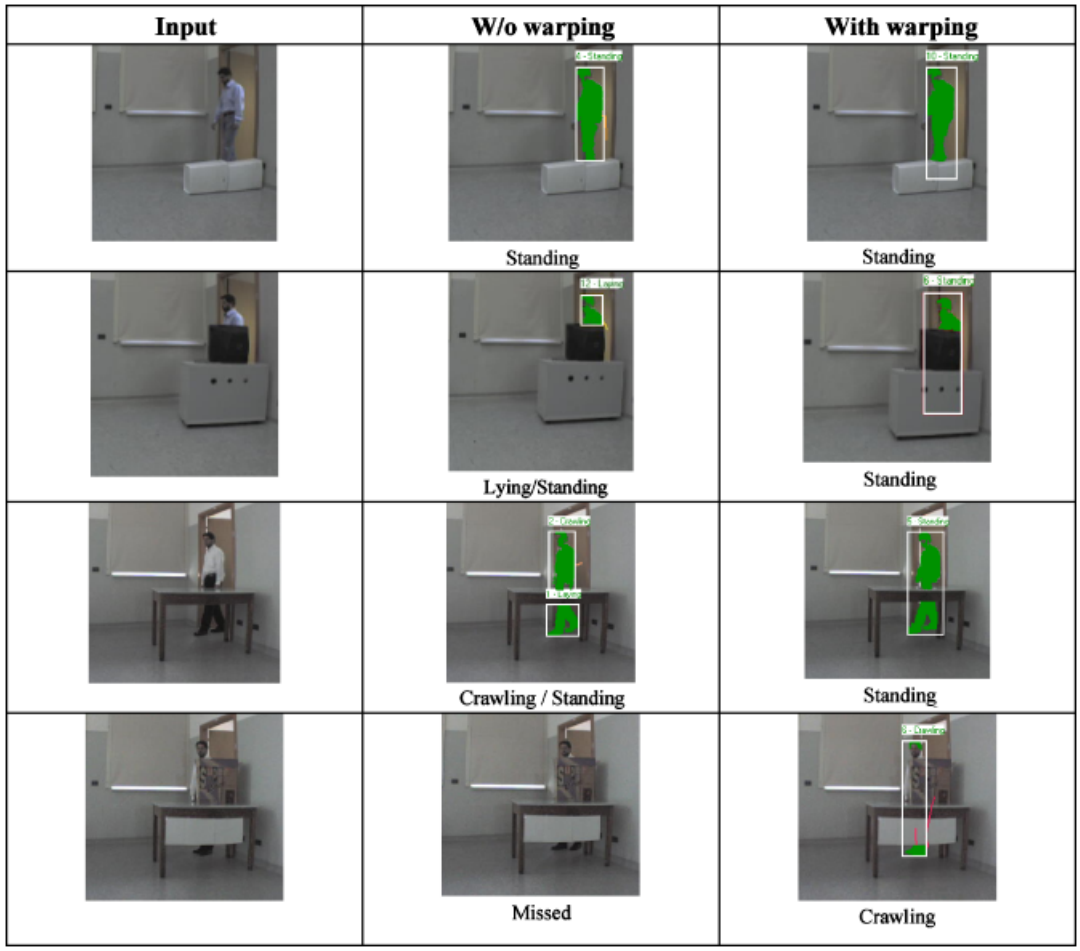


Figure 8: *Classification results with and without the multi-camera track warping.*

using track warping. In the second case (middle occlusions) the feet of the person are visible, but two or more tracks are generated and both the tracking and the posture classifier are misled.

Figure 8 reports the corresponding posture detection results, showing the classifications given by the system for all the frames subjected to the occlusion and a single snapshot as visual example. Whenever two postures are listed, this means that different postures are associated to either the same track in successive moments or two split tracks. Nevertheless, when the occluded part is too large, the warped track could be very different with respect to the segmented blob and the tracking algorithm is not capable of taking advantage of the initialization provided by the multi-camera module. As a consequence, after some frames the posture classifier fails (see the last row of Figure 8).

References

CAI, Q. and J.K. AGGARWAL (1999) Tracking human motion in structured environments using a distributed-camera system, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21** (12), 1241–1247.

CUCCHIARA, R., C. GRANA and A. PRATI (2003a) Semantic video transcoding using classes of relevance, *International Journal of Image and Graphics*, **3** (1), 145–169.

CUCCHIARA, R., C. GRANA, M. PICCARDI and A. PRATI (2003b) Detecting moving objects, ghosts and shadows in video streams, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25** (10), 1337–1342.

CUCCHIARA, R., C. GRANA, A. PRATI and R. VEZZANI (2003c) Computer vision techniques for PDA accessibility of in-house video surveillance, in *Proceedings of ACM Multimedia 2003 – First ACM International Workshop on Video Surveillance*, New York: ACM, 87–97.

CUCCHIARA, R., C. GRANA, A. PRATI and R. VEZZANI (2005) Probabilistic posture classification for human behaviour analysis, *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, **35** (1), 42–54.

HARITAOGLU, I., D. HARWOOD and L.S. DAVIS (1998) Ghost: a human body part labeling system using silhouettes, in *Proceedings of the International Conference on Pattern Recognition*, Brisbane: IEEE, 77–82.

HU, W., T. TAN, L. WANG and S. MAYBANK (2004) A survey on visual surveillance of object motion and behaviors, *IEEE Transactions on Systems, Man, and Cybernetics – Part C*, **34** (3), 334–352.

KANG, J., I. COHEN and G. MEDIONI (2003) Continuous tracking within and across camera streams, in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, New York: IEEE, Vol. 1, pp. 267–272.

KHAN, S. and M. SHAH (2003) Consistent labeling of tracked objects in multiple cameras with overlapping fields of view, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25** (10), 1355–1360.

LI, J., C.S. CHUA and Y.K. HO (2002) Color based multiple people tracking, in *Proceedings of the IEEE International Conference on Control, Automation, Robotics and Vision*, New York: IEEE, Vol. 1, pp. 309–314.

MOESLUND, T.B. and E. GRANUM (2001) A survey of computer vision-based human motion capture, *Computer Vision and Image Understanding*, **81**, 231–268.

RABINER, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, **77** (2), 257–286.

SENIOR, A. (2002) Tracking people with probabilistic appearance models, in *Proceedings of the International Workshop on Performance Evaluation of Tracking and Surveillance Systems*, Prague, 48–55.

YUE, Z., S.K. ZHOU and R. CHELLAPPA (2004) Robust two-camera tracking using homography, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, New York: IEEE, Vol. 3, pp. 1–4.

The authors

Rita Cucchiara

Rita Cucchiara (1989, Laurea in Electronic Engineering at the University of Bologna, Italy; 1993, PhD in Computer Engineering at the University of Bologna, Italy) is full Professor in the Faculty of Engineering at the University of Modena and Reggio Emilia. She is coordinator of the PhD curricula in computer science and engineering of the Modena ICT School of Doctorate, and heads the ImageLab Laboratory in the Dipartimento di Ingegneria dell'Informazione di Modena (http://imagelab.ing.unimo.it). Rita Cucchiara is responsible for many Italian and international projects. In European projects of the Sixth Framework Programme, she coor-

dinates the Unimore research unit in the VidiVideo project (2007–09) and DELOS, Network of Excellence in the Digital Library. She has been General Chair of the ACM Workshop of Video Surveillance and Sensor Networks (Singapore VSSN05, Santa Barbara VSSN06); in 2006 she was Guest Editor of the Special Issue on Multimedia Surveillance of the *ACM Journal of Multimedia*; and in 2007 she was General Chair of the International Conference on Image Analysis and Processing, Modena, Italy. She is on the Editorial Board of *Machine Vision and Applications* and *Multimedia Tools and Applications*. In 2006 she received the Fellowship of the International Association for Pattern Recognition for contributions in pattern recognition for video surveillance.

Andrea Prati

Andrea Prati is Assistant Professor in the Faculty of Engineering at Reggio Emilia, Italy. He received a PhD in Information Engineering at the University of Modena and Reggio Emilia. His research interests belong basically to three fields: performance analysis of multimedia computer architectures, motion analysis for surveillance applications, and semantic video transcoding. He collaborates in research projects at regional, national and European level. Andrea Prati is author or co-author of more than 70 papers in national and international journals and conference proceedings; he has been invited speaker, organizer of workshops and journals' special issues, and reviewer for many international journals in the field of computer vision and multimedia. He is a member of the IEEE, ACM and GIRPR (IAPR, Italy).

Roberto Vezzani

Roberto Vezzani is a post-doctoral research worker in the Faculty of Engineering at the University of Modena and Reggio Emilia. He received his PhD in Information Engineering from the University of Modena and Reggio Emilia in 2007. His research interests belong basically to automatic surveillance systems, covering problems of moving object detection, people tracking, surveillance data management and automatic annotation.