Making the home safer and more secure through visual surveillance

R. Cucchiara[1], A. Prati[2], R. Vezzani[1]
[1] *Dipartimento di Ingegneria dell'Informazione, University of Modena and Reggio Emilia, Italy*
[2] *Dipartimento di Scienze e Metodi dell'Ingegneria, University of Modena and Reggio Emilia, Italy*

Abstract
Video surveillance has a direct application in intelligent home automation or domotics (from the Latin word domus, that means "home", and informatics). In particular, in-house video surveillance can provide good support for people with some difficulties (e.g. elderly or disabled people) living alone and with limited autonomy. A key aspect in video surveillance systems for domotics is that of analyzing behaviours of the monitored people. To accomplish this task, people must be detected and tracked, and their posture must be analyzed in order to model behaviours recognizing abrupt changes in it.
Problems related to reliable software solutions are not completely solved, in particular luminance changes, shadows and frequent posture changes must be taken into account. Long-lasting occlusions are common due to the proximity of the cameras and the presence of furniture and doors that can often hide parts of a person's body. For these reasons, a probabilistic and appearance-based tracking, particularly conceivable for people tracking and posture classification, has been developed. However, despite its effectiveness for long-lasting and large occlusions, this approach tends to fail whenever the person is monitored with multiple cameras and he appears in one of them already occluded. Different views provided by multiple cameras can be exploited to solve occlusions by warping known object appearance into the occluded view. To this aim, this paper describes an approach to posture classification based on projection histograms, reinforced by HMM for assuring temporal coherence of the posture.

Keywords
People's behaviour analysis, computer vision, video surveillance, posture analysis

1 Introduction and Related Work
A very relevant application for behaviour analysis is that related to people's health care. This has been a crucial topic in many fields of research from almost the beginning. The increase each year in deaths and injuries for domestic incidents has shown up in-house safety as an emergent field of research. The widespread distribution of cameras in our world for video-surveillance purposes makes available a huge amount of visual information with which the people's safety can be improved.
With these premises, many computer vision techniques have been proposed to (semi-)automatically assure the people's safety. Among them, the detection of people's posture has recently gained credit, because of fundamental importance for people's behaviour analysis or for detecting alarming situations (such as a fall). 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 2D or 3D model of the human body [9]. 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 in [4]) 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 behaviours 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 solutions are preferable, thus stereovision or 3D 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 [4, 9].
Since the posture is often related to the shape, occlusions are very critical. Even though some probabilistic approaches are able to maintain the shape in the case of occlusions (as in [4]) they are likely to fail if the occlusion occurs at the beginning, i.e. when the track is firstly 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 to have a full coverage of the environment and to solve occlusions does not come for free. The technical problems in multiple camera systems are several and they have been summarized in [5] into six classes: installation, calibration, object matching, switching, data fusion, and occlusion handling. Among these, object matching is the most addressed problem in the literature and provides the basic tools also for the occlusion handling.
Several works have been proposed to maintain the correspondence of the same tracked object during a camera handoff. Most of those require a partially overlapping field of view [1]; other ones use a feature based probabilistic framework to maintain a coherent labelling of the objects [7]. All these works aim at keeping correspondences between the same tracked object (in order to continuously analyze the behaviour in a wide area), but none of them are capable of handling the 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 of calibrated approach is reported in [12] in which homography is exploited to solve occlusions. Single camera processing is based on particle filter 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 very relevant example of the uncalibrated approaches is the work of Khan and Shah [7]. Their approach 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. 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.

Colour-based approaches base the matching essentially on the colour of the tracks, as in [8] 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 [6] or Bayesian Belief Networks (BBN) [1].

Our approach is similar to that proposed in [12], but, differently from it, appearance models of the tracks are warped from one view to another not using 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 experimentation that aims at analyzing the limits of the approach depending on the amount and type of the occurred occlusion.

2 Single Camera Behaviour Analyzer

In our multi-camera system, moving objects are extracted from each camera by exploiting background suppression with selective and adaptive update in order to react quickly to the changes and to also take "ghosts" (i.e., aura left behind by an object that begins to move) into account [2]. 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 in [11], is used to handle objects with non rigid motion, variable shape (like people), and frequent occlusions. This tracking algorithm maintains, in addition to the current blob *B*, the appearance image *AI* (or temporal template) and the probability mask *PM* of the track. *AI* is obtained with a temporal integration of the color images of the blobs, while the probability mask *PM* associates to 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 [3].

Finally, tracks that satisfy some geometrical and color constraints are classified as people and submitted to the posture classifier. Four main postures are considered: *Standing, Crawling, Sitting,* and *Lying*. To this aim, similarly to [4], we exploit a classifier based on the projection histograms computed over the blobs of the segmented people. The projection histograms $PH = (\vartheta(x); \pi(y))$ 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, consequentially, on the position of the person inside the room, we first scale them according with the distance of the person with respect to the camera. To compute this distance, a feet detection and tracking module together with a homography relation obtained through camera calibration are exploited [3].

Though 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. To this aim, a Hidden Markov Model formulation has been adopted. Using the notation proposed by Rabiner in his tutorial [10], the followings sets are defined:

- the state set S, composed by N states: $S = \{S_1, ..., S_N\} = Main\_Postures$
- 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 **A** of the state transition probabilities, computed as a function of a reactivity parameter α (empirically determined; for example, we set α = 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 **A** has the following structure:

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

The Observation Symbols and the Observation Symbol Probability distribution **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 by *N* rows and w^h columns (where *w* and *h* are the sizes of the images). Thus, we prefer to directly compute the probability values b_j, that indicate the probability to have a particular observation (histograms) belonging to the state (posture) *j*, through the output of the frame-by-frame classifier:

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

The HMM presented does not require any additional training phase because it exploits directly the Probability Maps. Then, at each frame, the probability of being in each state is computed with the traditional forward algorithm. At last, the HMM input has been modified to keep into account the visibility status of the person. In fact, if the person is completely occluded, the reliability of the posture must decrease with the time. In such a situation, it is preferable to set $b_j = 1/N$ as the input of the HMM. Making that, 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 to be in a state S_j, the higher the time required to lose this certainty. To manage simultaneously the two situations and to cope with the intermediate cases, (i.e., partial occlusions), a generalized formulation of the HMM input is defined:

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

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 Eq. 2, otherwise the higher the value of n_{fo}, the more it tends to a uniform distribution.

In Figure 1, the benefits of the introduction of the HMM framework are evidenced. 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 laying). Instead, through the HMM framework, the temporal coherence of the posture is preserved, even if the classification reliability decreases during the occlusion.

Posture classification is the first, essential step for behaviour analysis. For instance, dangerous situations such as falls, can be identified by detecting abrupt changes from a standing to a laying posture. More advanced behaviours can be detected and analyzed by including also other mid-level features, such as the people's gait, into the modelling of the behaviours.

3 Multi-camera Occlusion Manager

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 in the case the track is occluded since its creation (as in Figure 2.b).

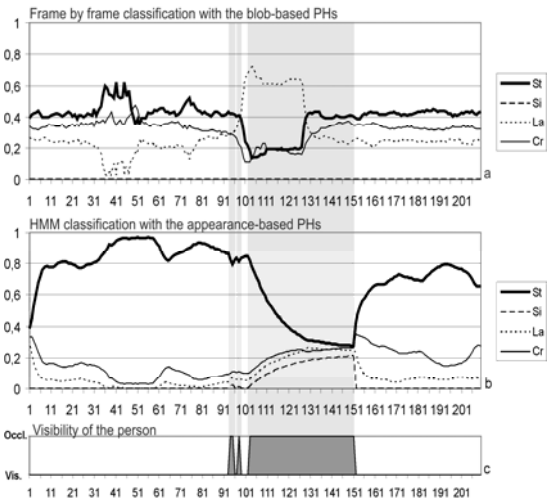


Figure 1. *Frame by frame and HMM posture classification during a strong occlusion*

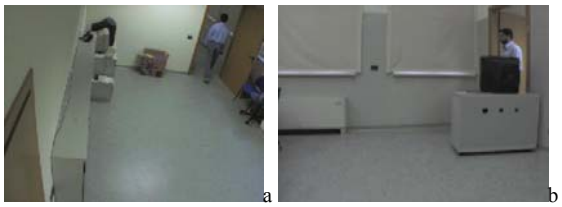


Figure 2. *An example of track occlusion during its creation: the source (a) and destination view (b)*

Our proposal is to exploit the appearance information from another camera (where the track is not occluded) to solve this problem. If a person passes between two monitored rooms, it is possible to keep the temporal information stored into its track extracted from the first room (Figure 2.a) and use them to initialize the corresponding track in the second room (Figure 2.b).

To this aim, with reference to Figure 3, we assume that 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 when the person passes into the second room; all the track points lie on a plane P parallel to the entrance plane G and containing the feet.

The first three assumptions imply only an accurate installation and calibration of cameras and sensors, while the last one is a simplification needful to warp the track between the two points of view. Under this condition, in fact, the 3D 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.

In particular, the above mentioned process is applied only to the four corners of the tracks, and, exploiting them, the homography matrix H that transforms each point between the two views can be computed. Through H it is possible to re-project both the appearance image AI and the probability mask PM of each track from the point of view of the leaving room to the point of view of the entering one. 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.

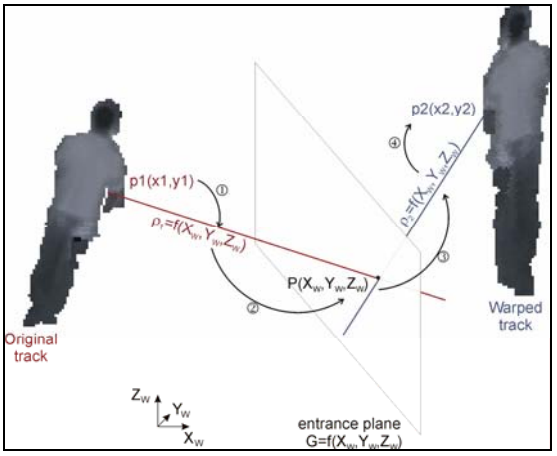


Figure 3. *Track warping: 1) exploiting the calibration of camera1, 2) intersection with entrance plane, 3) calibration of camera2, 4) intersection with camera2 plane*

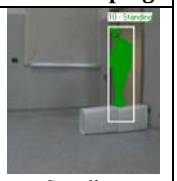
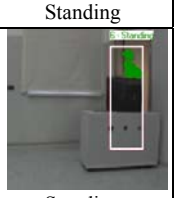
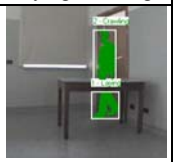
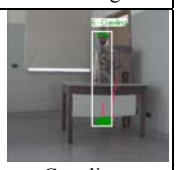
Input	W/o warping	With warping
	 Standing	 Standing
	 Laying/Standing	 Standing
	 Crawl. / Standing	 Standing
	 Missed	 Crawling

Table 1. *Classification results with and without the multicamera track warping*

4 Experiments and Conclusions

As a test bed for our multi-camera system, a two-rooms setup has been created. The two rooms share a door equipped by an optical sensor used to trigger the passage of the 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. 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.

Table 1 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 last row of Table 1).

References:

1. Cai, Q.; Aggarwal, J.K. (1999). Tracking Human Motion in Structured Environments Using a Distributed-Camera System. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(**12**),1241-1247.
2. Cucchiara, R.; Grana, C.; Piccardi, M.; Prati, A. (2003). Detecting Moving Objects, Ghosts and Shadows in Video Streams". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(**10**):1337-1342.
3. Cucchiara, R.; Grana, C.; Prati, A.; Vezzani, R.. (2005). Probabilistic Posture Classification for Human Behaviour Analysis. *IEEE Trans. on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 35(**1**), 42-54.
4. Haritaoglu, I.; Harwood, D.; Davis, L.S.. (1998). Ghost: A Human Body Part Labeling System Using Silhouettes. P*roc. of Intl Conf. on Pattern Recognition.*
5. Hu, W.; Tan, T.; Wang, L.; Maybank, S. (2004). A Survey on Visual Surveillance of Object Motion and Behaviours. *IEEE Trans. on Systems, Man, and Cybernetics - Part C*, 34(**3**):334–352.
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. Khan, S.; Shah, M. (2003). Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(**10**):1355-1360.
8. Li, J.; Chua, C.S.; Ho, Y.K (2002). Colour Based Multiple People Tracking. *Proc. of IEEE Intl Conf. on Control, Automation, Robotics and Vision*, **1**, 309–314.
9. Moeslund, T.B.; Granum, E, (2001). A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, **81**, 231-268.
10. Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(**2**):257–286.
11. Senior, A. (2002). Tracking People with Probabilistic Appearance Models. *Proc. of Intl Workshop on Performance Evaluation of Tracking and Surveillance Systems*.
12. Yue, Z.; Zhou, S.K.; Chellappa, R. (2004). Robust Two-camera Tracking using Homography. *Proc. of IEEE Intl Conf. on Acoustics, Speech, and Signal Processing*, **3**, 1–4.