A DISTRIBUTED OUTDOOR VIDEO SURVEILLANCE SYSTEM FOR DETECTION OF ABNORMAL PEOPLE TRAJECTORIES

S. Calderara, R. Cucchiara

Univ. of Modena and Reggio Emilia
D.I.I. - Italy

A. Prati

Univ. of Modena and Reggio Emilia
Di.S.M.I. - Italy

ABSTRACT

Distributed surveillance systems are nowadays widely adopted to monitor large areas for security purposes. In this paper, we present a complete multicamera system designed for people tracking from multiple partially overlapped views and capable of inferring and detecting abnormal people trajectories. Detection and tracking are performed by means of background suppression and an appearance-based probabilistic approach. Objects' label ambiguities are geometrically solved and the concept of "normality" is learned from data using a robust statistical model based on Von Mises distributions. Abnormal trajectories are detected using a first-order Bayesian network and, for each abnormal event, the appearance of the subject from each view is logged. Experiments demonstrate that our system can process with real-time performance up to three cameras simultaneously in an unsupervised setup and under varying environmental conditions.

Index Terms— One, two, three, four, five

1. INTRODUCTION

It is undoubted that distributed video surveillance is a hot topic of research. Finding a way to efficiently exploit the hundreds of cameras installed in our cities is a must in order to reduce the required "human effort" necessary to monitor such huge number of cameras and to automatically extract useful information from them. Computer vision can help in accomplishing to this difficult task.

Having distributed cameras, the first key task is to detect and track moving objects (especially people) in each camera. Computer vision approaches for people detection and tracking have been deeply studied in the last decade. Indoor and outdoor applications for video surveillance have been developed both in research projects (such as Pfinder [1], VSAM [2] or W4 [3]) and, more recently, in commercial systems. Moreover, since in real scenarios the scene is cluttered and complex, occlusions and people overlaps suggested the adoption of probabilistic techniques to allow robust tracking of the person's position [4, 5].

Once people are tracked on each camera, it can be convenient to try to have consistent labels/identities of the same person/object when it moves in different views. This is known as "the consistent labeling problem" and it has been studied independently by the adopted tracking method. Three different approaches are clearly identifiable in the literature. The former bases the matching essentially on the color of the objects [6]. Although these techniques are error-prone, due to the strict dependency on the environmental condition, are the only feasible when the cameras' fields of view do not overlap. The second class of approaches exploits geometrical information only to solve label ambiguities. Exploiting calibration, the relationship between overlapped cameras can be easily modelled in the 3D space and warping techniques can be applied with high accuracy [7]. Conversely, a fully uncalibrated approach, based on the image projections of overlapped cameras' field of view lines, has been initially proposed by Khan and Shah in [8]. The latter class combines information about the geometry with information provided by the visual appearance. Different techniques are adopted to fuse information, based on probabilistic information fusion [9] or on Bayesian Belief Networks (BBN) [10].

Finally, several papers (e.g. [11]) have been proposed for trajectory classification, but none of them is based on modeling abnormal paths.

2. SYSTEM OVERVIEW

The presented system is a complex set of modules depicted in the scheme of Fig. 1. The main scope of this paper is to report the overview of the whole system and briefly describe each single module. Readers interested to more details on these modules can refer to the cited papers.

The acquisition platform is composed of several cameras, both fixed and PTZ. Each of the modules connected with a fixed camera segments and tracks moving people.

Moving people segmentation is achieved by using the background suppression approach called SAKBOT (Statistical And Knowledge-Based Object Tracker) presented in [12]. Proper techniques for background bootstrapping, ghost suppression, and object validation have also been introduced to improve the accuracy of the segmentation in cluttered outdoor setup (see Section 3).

1-4244-1354-0/07/$25.00

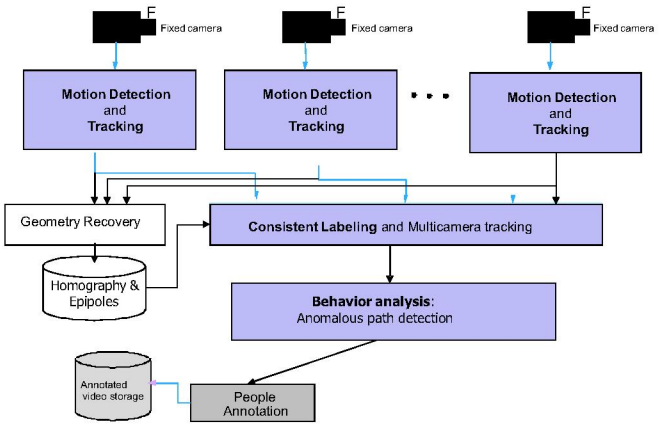


Fig. 1. Diagram of the outdoor surveillance system developed at Imagelab.

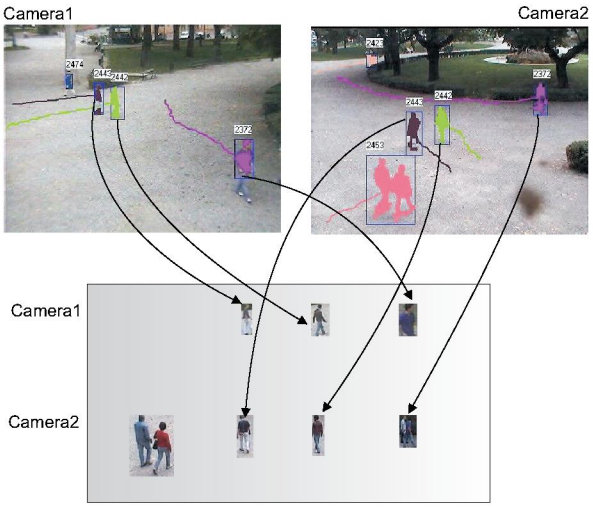


Fig. 2. Example of people logging.

Moving objects detected by SAKBOT are then classified in person or not-person according with their geometrical shape and size (scaled taking into account their position on the ground plane) and then tracked in each single view by means of an appearance-based algorithm [13] (see Section 3).

As shown in Fig. 1, the outputs of each camera module are processed by the Bayesian-competitive consistent labeling module. This module aims at assigning the same label to the different instances of the same person viewed by different cameras with overlapped fields of view (see Section 4).

The global label assignment assured by the consistent labeling is the fundamental step for the subsequent, higher-level tasks shown in Fig. 1. For instance, the information provided by the multi-camera tracking system can be further analyzed to detect anomalous people paths in the scene. This task is accomplished by learning the concept of "normality" modeling statistically the paths recurrence as a composition of Von Mises probability density functions (see Section 5). For each anomalous detected event is then possible to store multiple snapshots from multiple viewpoints of the involved person that can be used for successive retrieval of information (see Fig. 2).

3. MOTION DETECTION AND TRACKING

The adopted motion detection algorithm is specifically design to ensure a robust and reliable background estimation even in complex outdoor scenarios. It is a modification of the SAKBOT system [12], that increases the robustness in outdoor uncontrolled environment. The SAKBOT background model is a temporal median model with a selective knowledge-based update stage. Suitable modification to the background initialization, motion detection and object validation have been developed. The initial background model at time t, BG, is initialized subdividing the input image I in 16×16 size blocks. For each block, a single difference over time, with the input frame I_t, is performed and the number of still pixels are accounted as the block's weight. The background is then selectively updated by including all the blocks composed by more than 95% of still pixels and the initialization process halts when the whole background image BG is filled with "stable" blocks.

After the bootstrapping stage, the background model is updated using a selective temporal median. A fixed k-sized circular buffer is used to collect values of each pixel over time. In addition to the k values, the current background model $BG_t(i,j)$ is sampled and added to the buffer to account for the last reliable background information available. These $n = k + 1$ values are then ordered according to their gray-level intensity, and the median value is used as an estimate for the current background model.

The difference between the current image I_t and the background model BG is computed and then binarized using two different local and pixel-varying thresholds: a low threshold T_{low} to filter out the noisy pixel extracted due to small intensity variations; a high threshold T_{high} to identify the pixels where a large intensity variation occurs:

$$T_{low}(i,j) = \lambda \left(b_{\frac{k+1}{2}+l} - b_{\frac{k+1}{2}-l} \right) \quad (1)$$

$$T_{high}(i,j) = \lambda \left(b_{\frac{k+1}{2}+h} - b_{\frac{k+1}{2}-h} \right) \quad (2)$$

where b_p be the value at position p inside the ordered circular buffer b of pixel (i,j) and λ, l and h are fixed scalar values.We experimentally set $\lambda = 7$, $l = 2$ and $h = 4$, for a buffer of $n = 9$ values. The final binarized motion mask M_t is obtained as composition of the two binarized motion masks computed respectively using the low and the high thresholds: a pixel is marked as foreground in M_t if it is presented in the low-thresholded binarized mask AND it is spatially connected to at least one pixel present in the high-thresholded binarized mask.

Finally, the list MVO_t of moving objects at time t is extracted from M_t. Objects are then validated using jointly color, shape and gradient information to remove artifacts and objects due to small background variations and invalid objects are directly injected in the background model (see [14] for further details). Shadows detection is also performed in the HSV color space [15]. Shadow objects are then discarded and do not concur to the background model update stage.

One of the problem of selective background updating is the possible creation of ghosts. The approach used to detect and remove ghosts is similar to that used for background initialization, but at region level instead of pixel level. All the validated objects are used to build an image called $A_t(i,j)$ that accounts for the number of times that a pixel is detected as stopped by the single difference.

A valid object MVO_t^h is classified as ghost if:

$$\frac{\sum_{(i,j)\in MVO_t^h} A_t(i,j)}{N_t^h} > T_{ghost} \tag{3}$$

where T_{ghost} is the threshold on the percentage of points of the MVO_t^h stopped for sufficient time.

The objects detected as moving and validated as people are then tracked in each single view by means of an appearance-based algorithm. The algorithm uses a classical predict-update approach. It takes into account not only the status vector containing position and speed, but also the memory appearance model and the probabilistic mask of the shape [13]. The former, also called dynamic template in [3], is the adaptive update of each pixel in the color space. The latter is a mask whose values in the range between 0 and 1 can be viewed as the probability for that pixel to belong to that object. These models are used to define a MAP (Maximum A Posteriori) classifier that searches the most probable position of each person in the scene. The tracking algorithm includes a specific module for coping with large and long-lasting occlusions. The occlusion handling is very robust and has been tested in many applications. It can keep the shape of the tracked objects very precisely and has been exploited for both outdoor and indoor video-surveillance and in different applications such as people posture classification [16].

4. BAYESIAN-COMPETITIVE CONSISTENT LABELING

Acquiring data from multiple video streams lead to the problem of mantaining the person identity consistent among cameras. This problem is solved adopting a geometric approach that exploits cameras' FoV relations and constraints to impose identities' consistency that we called HECOL (Homography and Epipolar-based COnsistent Labeling). In detail, when cameras partially overlap, the shared portion of the scene is analyzed and people identities are matched geometrically. After an initial unsupervised and automatic training phase the overlapping regions among field of views, ground-plane homographies and the epipole location for pairwise overlapping cameras are computed. The consistent labeling problem is then solved on-line whenever a new object τ appears in the field of view of a given camera C^1, where superscript indicates the camera ID. The multi-camera system must check whether τ corresponds to a completely new object or to one which is already present in the FOV of other cameras. Moreover, the system should deal with groups and identify the objects composing them. The presented approach has the advantage of coping with labeling errors and partial occlusions whenever the involved objects are present in at least one overlapped view. Using the vertical objects' inertial axis as discriminant feature can also help to disambiguate the group of people detected as a single blob exploiting the information on an overlapped views.

When many objects are present in the scene and many cameras are involved, the exhaustive search may be computationally expensive. Thus, the subset of K potential matching objects satisfying the camera topology constraints are efficiently extracted by means of a graph model (called *Camera Transition Graph*). These K objects are combined to form the hypothesis space Γ that contains all the $\left(2^K-1\right)$ possible matching hypotheses, with both single object and groups. A MAP estimator is adopted to find the most probable hypothesis $\gamma_i \in \Gamma$ being:

$$i = \arg\max_k \left(p\left(\gamma_k \mid \tau\right)\right) = \arg\max_k \left(p\left(\tau \mid \gamma_k\right) p\left(\gamma_k\right)\right) \tag{4}$$

To evaluate the maximum posteriori the prior of each hypothesis γ_k and the likelihood of the new object τ given the hypothesis must be computed. The prior of a given hypothesis γ_k is not computed by means of a specific pdf, but it is heuristically evaluated by assigning a value proportional to a score σ_k. The score σ_k accounts for the distance between objects calculated after the homographic warping. A hypothesis consisting of a single object will then gain higher prior if the warped lower support point (i.e. the point of the object that contacts with the ground plane) $\mathbf{lp}$ is far enough from the other objects' support points. On the other hand, a hypothesis consisting of two or more objects (i.e., a possible group) will gain higher prior if the objects that compose the hypothesis are close to each other after the warping, and, at the same time, the whole group is far from other objects.

Let us suppose the new object τ appears on camera C^1. The $\mathbf{lp}$ of each of the K objects in C^2 is warped to the image plane of C^1. Likelihood is then computed by testing the fitness of each hypothesis against current evidence. The main goal is to distinguish between single hypothesis, group hypotheses and possible segmentation errors exploiting only geometrical properties in order to avoid the uncertainties due to color variation and adopting the vertical axis of the object as an invariant feature.

The axis of the object τ can be warped correctly only with the homography matrix and the knowledge of epipolar con-

straints among cameras. To obtain the correct axis inclination the vertical vanishing point (computed by a robust technique as described in [17]) is then used as shown in Figure 3. The lower support point $\mathbf{lp}$ of τ is projected on camera C^2 by using the homography matrix. The corresponding point on the image plane of camera C^2 is denoted as $\mathbf{a_1} = H\mathbf{lp}$, where H is the homography matrix from C^1 to C^2. The warped axis will lie on a straight line passing through $\mathbf{vp^2}$ and $\mathbf{a_1}$ (Fig. 3(d)). The ending point of the warped axis is computed by using the upper support point $\mathbf{up}$ that is the middle point of the upper side of the object's bounding box. Since this point does not lie on the ground plane, its projection on the image of camera C^2 does not correspond to the actual upper support point; however, the projected point lies on the epipolar line. Consequently, the axis' ending point $\mathbf{a_2}$ is obtained as the intersection between the epipolar line $\langle \mathbf{e^2}, H\mathbf{up} \rangle$ and line $\langle \mathbf{vp^2}, H\mathbf{lp} \rangle$ passing through the axis.

Based on geometrical constraints, the warped axis $\langle \mathbf{a_1}, \mathbf{a_2} \rangle$ of τ in the image plane of C^2 is univocally identified but its computation is not error free. In order to improve the robustness to computation errors, we account also for the dual process that can be performed for each of the K potential matching objects: the axis of the object in C^2 is warped on the segment $\langle \mathbf{a_1}, \mathbf{a_2} \rangle$ on camera C^1.

The measure of axis correspondence is not merely the distance between axes $\langle \mathbf{a_1}, \mathbf{a_2} \rangle$ and $\langle \mathbf{lp}, \mathbf{up} \rangle$; it is defined as the number of matching pixels between the warped axis and the foreground blob of the target object. This makes it easier to define a normalised value for quantifying the matching. Accordingly, the fitness measure $\varphi_{\tau_a \to \tau_b}$ from the object τ_a in a generic camera C^i to τ_b in a generic camera C^j is defined as the number of pixels resulting from the intersection between the warped axis and the foreground blob of τ_b normalized by the length (in pixels) of the warped axis itself. The reversed fitness measure $\varphi_{\tau_b \to \tau_a}$ is computed similarly by reverting the warping order. In the ideal case of correspondence between τ_a and τ_b, $\varphi_{\tau_a \to \tau_b} = \varphi_{\tau_b \to \tau_a} = 1$. However, in the case of errors in the $\mathbf{lp}$ and $\mathbf{up}$ computation, the warped axis could fall partially outside of the foreground blob, lowering the fitness measure.

In the likelihood definition we refer to *forward* contribution when the fitness is calculated from the image plane in which the new object appears (camera C^1) to the image plane of the considered hypothesis (camera C^2). Thus, generalizing for hypotheses containing more than one object (group hypotheses), forward axis correspondence can be evaluated by computing the fitness of the new object τ with all the objects composing the given hypothesis γ_k for camera C^2:

$$fp_{forward}(\tau|\gamma_k) = \frac{\sum_{\tau_m \in \gamma_k} \varphi_{\tau \to \tau_m}}{K \cdot S_f} \tag{5}$$

S_f measures the maximum range of variability of the forward fitness measure of the objects inside the given hypothesis:

$$S_f = \max_{\tau_m \in \gamma_k} (\varphi_{\tau \to \tau_m}) - \min_{\tau_n \in \gamma_k} (\varphi_{\tau \to \tau_n}) \tag{6}$$

The use of the normalizing factor K (i.e. the number of potential matching objects on C^2) weighs each hypothesis according to the presence or absence of objects in the whole scene.

Backward contribution is computed similarly from the hypotheses space to the observed object:

$$fp_{backward}(\tau|\gamma_k) = \frac{\sum_{\tau_m \in \gamma_k} \varphi_{\tau_m \to \tau}}{K \cdot S_b} \tag{7}$$

where S_b is defined as:

$$S_b = \max_{\tau_m \in \gamma_k} (\varphi_{\tau_m \to \tau}) - \min_{\tau_n \in \gamma_k} (\varphi_{\tau_n \to \tau}) \tag{8}$$

At the end, the likelihood is defined as the maximum value between forward and backward contribution. The use of the maximum value ensures that the contribution where the extraction of support points is generally more accurate and suitable for the matching will be used. The effectiveness of the double backward/forward contribution is evident in the full characterization of groups of people. The forward contribution helps to solve the situations when a group of objects is already inside the scene while the group's components appear one at a time in another camera. On the contrary the backward component is useful when two people appear in a new camera detected as a single blob. The group disambiguation can be solved by exploiting the fact that in the other camera the two objects are detected as separated. Backward contribution is also useful to solve the case of *segmentation errors*, in which a person has been erroneously extracted by the object detection system as two separate objects, but a full view of the person exists from the past in an overlapped camera. The results reported in Section 6 will demonstrate this.

When more than two cameras overlap simultaneously it is possible to take into account more information than in the pairwise case. To account for this situation the proposed approach is suitably modified by adding an additional step that selects the best assignment from all the possible hypotheses coming from each camera. In details, when a detection event occurs on C^1, for each camera C^j overlapped with C^1 the best local assignment hypothesis is chosen using the maximum-a-posteriori framework. A second MAP estimator detects the most probable among these hypotheses. In complex scenes more hypotheses could have similar a-posteriori probability but it may exist a particular view where the hypothesis assignment is easier. The second MAP stage has the purpose to choose this view; this can be easily done using the previously computed posteriors and Bayes rule:

$$p(C^j \mid \tau) \propto p(\tau \mid C^j) = \max_{\gamma_k \in \Gamma} p(\gamma_k \mid \tau) \tag{9}$$

The camera posterior is evaluated for each camera C^j that overlaps with C^1 assuming that all the views of overlapped

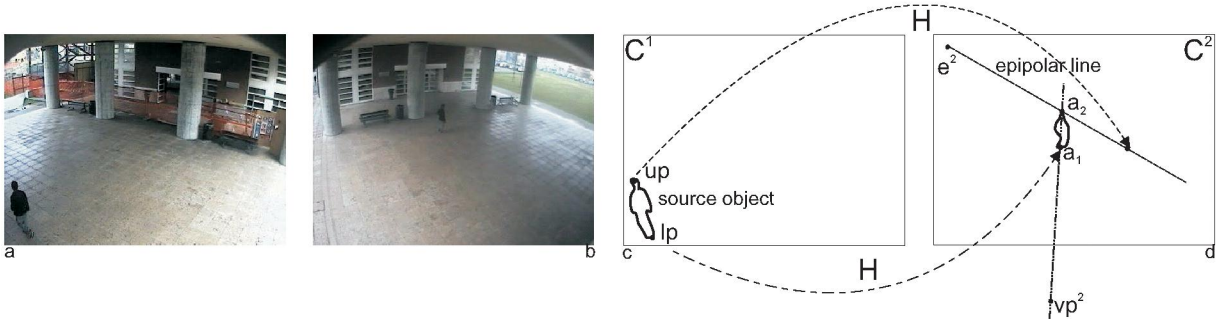


Fig. 3. Example of exploiting vanishing point and epipolar geometry to warp the axis of the object τ to the image plane of camera C^1.

cameras are equally probable. Eventually, the label is assigned to the new object according to the winning hypothesis on the winning camera. If the chosen hypothesis identifies a group, all the labels of objects composing the group are assigned as identifiers.

5. PEOPLE PATH CLASSIFICATION

After the label disambiguation process, the tracking output can be exploited for further high-level reasoning on people behavior. In particular paths can be analyzed to detect whenever anomalous events occur in the system. People paths are extracted directly from the multi-camera tracking system and homographically projected onto the ground plane. Each path is modeled as a sequence of directions computed as the angle between two consecutive points. Due to this approximation of the direction, we applied a running average filter of fixed size to smooth the segmentation errors and discretization effects on the direction computation. Each of these paths is then modeled with a Von Mises distribution [18] [19]. This distribution is also called "circular normal" since it is equivalent to a Gaussian/normal distribution but applied to a periodic variable. In fact, standard Gaussian distributions are not suitable for periodic variable, such as the direction. Treating periodic variable by setting a value as origin and then applying traditional Gaussian distribution will bring to results that were strongly dependent on the arbitrary choice of the origin. Conversely, Von Mises distribution is independent on the origin, being "circularly defined".

In our case, the periodic variable is the direction θ which has a period of 2π. Thus, the Von Mises distribution for the direction θ is defined as:

$$p\left(\theta|\theta_0, m\right) = \frac{1}{2\pi I_0\left(m\right)} e^{m\cos(\theta-\theta_0)} \tag{10}$$

where I_0 is the modified zero-order Bessel function of the first kind, defined as:

$$I_0\left(m\right) = \frac{1}{2\pi}\int_0^{2\pi} e^{m\cos\theta} d\theta \tag{11}$$

and represents the normalization factor. The parameters θ_0 and m represent, respectively, the mean and the precision (inverse of the variance) of the distribution.

Von Mises distribution has the limit to be unimodal, describing, in our case, the main direction θ_0 of a path and the associated range of variability by means of the precision m. We model the class of "normal" learned behavior as a composition of K Von Mises distributions to account for different main directions and to filter out outliers by means of clustering. Therefore, clustering is required to compute in a probabilistic way the best K representatives of such a composition, i.e. the K main possible directions of the "normal" path. Von Mises parameters are learned using a maximum likelyhood estimator under the hypothesis that the directions composing a path are i.i.d. Knowledge of the concept of normality is built considering all the observed path over time and clustering together similar distribution. Clustering the paths modeled by parameters of Von Mises distributions can not be done using Euclidean distance in the parameter space for this reason, it is more correct to use *Bhattacharyya* distance to compare two distributions [20]. Let $p\left(\theta\right)$ and $q\left(\theta\right)$ be two distributions on a variable θ, then we can compute the quantity:

$$\Omega\left(p, q\right) = \int_{-\infty}^{+\infty} \sqrt{p\left(\theta\right) q\left(\theta\right)} d\theta \tag{12}$$

The Bhattacharyya distance can be compute by either $\widetilde{B} = -\ln\Omega$ or $B = \sqrt{1-\Omega}$. Since the first expression is not actually a distance (hence called *similarity* more properly), i.e. does not satisfy the triangular inequality, the second expression is preferred. A closed form of the Bhattacharyya distance for Von Mises distribution has been proposed in [21]. By defining two Von Mises distributions $p\left(\theta\right)$ and $q\left(\theta\right)$ as:

$$\begin{aligned} p\left(\theta\right) &= p\left(\theta|\theta_{0,1}, m_1\right) = \frac{1}{2\pi I_0\left(m_1\right)} e^{m_1\cos(\theta-\theta_{0,1})} \\ q\left(\theta\right) &= q\left(\theta|\theta_{0,2}, m_2\right) = \frac{1}{2\pi I_0\left(m_2\right)} e^{m_2\cos(\theta-\theta_{0,2})} \end{aligned} \tag{13}$$

the Bhattacharyya distance can be obtained by the following

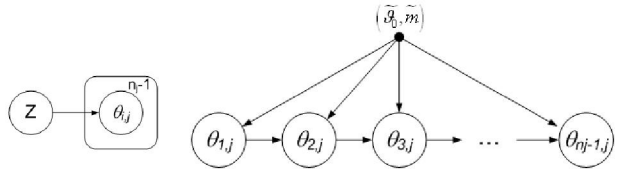
closed-form equation:

$$B(p,q) = \sqrt{1 - \left(\sqrt{\frac{1}{I_0(m_1)\, I_0(m_2)}}\, I_0\left(\frac{\sqrt{m_1^2 + m_2^2 + 2m_1 m_2 \cos(\theta_{0,1} - \theta_{0,2})}}{2}\right)\right)} \tag{14}$$

By means of the Bhattacharyya distance, the clustering of the parameter space can be performed by using *k-medoids* algorithm [22]. This is a suitable modification of the well-known k-means algorithm which has the appreciable characteristic to compute as prototype of the cluster the median (corresponding to an actual distribution) instead of the mean. In other words, at each iteration the prototype of each cluster is given by the member of the cluster at the minimum average distance from all the other members. Each cluster then gain a prior probability proportional to the number of elements composing it; this allows to have abnormal elements in the training set and to probabilistically filter out outliers.

5.1. Path classification using Bayesian Networks

After the training process, K medoids with the corresponding Von Mises distribution parameters are obtained. The classification of a new path T_j is performed by a two-steps approach that first selects the best candidate model among the available medoids and subsequently tests its fitness with the observed data. The model selection is obtained exploiting a Bayesian



(a) First step - original model (b) Second step - Bayesian network

Fig. 4. Probabilistic model used. (a) reports the statistical model used for the first step aiming at finding the best medoid to associated with the observed directions. (b) presents the first-order Bayesian network used in the second step. The rounded box in (a) represents a plate.

MAP framework that maximizes the model's parameters posterior over the observed path directions as shown in Fig. 4(a). To perform the inference a discrete hidden variable Z is introduced. Z is a 1-of-K variable and can be represented by a vector of binary values Z_i, one for each medoid, distributed as shown in the following equation:

$$p(Z) = \prod_{i=1}^{8} \mu_i^{Z_i} \tag{15}$$

where μ_i is equal to the prior probabilities of the i^{th} medoid, computed by considering the number of elements belonging to the cluster with respect to the total number of elements in the training set. By exploiting the Bayes rule, the posterior probability $p(Z_i|T_j)$ of a given model Z_i can be evaluated assuming that the directions of a given path are independent each other. This strong assumption will be further relaxed to test the model fitness.

After the evaluation, the desired model parameters are selected according to the Z_i value that maximizes the posterior:

$$p(Z_i|T_j) \propto \prod_{l=1}^{n_j-1} p(\theta_{l,j}|Z_i = 1)\; p(Z_i) \tag{16}$$

where $p(\theta_{l,j}|Z_i = 1)$ is a Von Mises distribution with the parameters of the i^{th} medoid and $p(Z_i)$ is equal to μ_i. The parameters of the selected medoid will be hereinafter denoted with $\widetilde{\theta}_0$ and $\widetilde{m}$.

After the selection of the model, the path T_j is verified against the model using the first-order Bayesian network shown in Fig. 4(b). By applying the D-separation property the probability $p\left(T_j|\widetilde{\theta}_0, \widetilde{m}\right)$ factorizes as:

$$p\left(T_j|\widetilde{\theta}_0, \widetilde{m}\right) = p\left(\theta_{1,j}|\widetilde{\theta}_0, \widetilde{m}\right) \prod_{l=2}^{n_j-1} p\left(\theta_{l,j}|\theta_{l-1,j}, \widetilde{\theta}_0, \widetilde{m}\right) \tag{17}$$

The conditional probability at right-hand side of Eq. 17 can be decoupled in two contributions, the first coming by the fitness of the variable $\theta_{l,j}$ (i.e., the given node of the Bayesian network) against the selected model and the second coming from the range of variability with respect to the the previous observed value.

The conditional probability $p\left(\theta_{l,j}|\theta_{l-1,j}, \widetilde{\theta}_0, \widetilde{m}\right)$ becomes:

$$p\left(\theta_{l,j}|\theta_{l-1,j}, \widetilde{\theta}_0, \widetilde{m}\right) = \frac{1}{2\pi I_0(\widetilde{m})} e^{\widetilde{m}\cos\left(\theta_{l,j} - \widetilde{\theta}_0\right)} \cdot \frac{1}{2\pi I_0(\widetilde{m})} e^{\widetilde{m}\cos(\theta_{l,j} - \theta_{l-1,j})} \tag{18}$$

Expliciting the product of two Von Mises we obtain a single Von mises distribution having parameters $\overline{\theta_0}$ and $\overline{m}$:

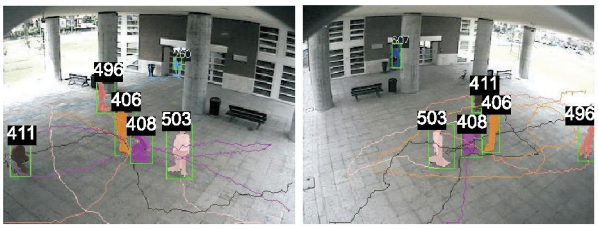
$$\begin{aligned} \overline{m} &= 2\widetilde{m}\cos\left(\frac{\widetilde{\theta}_0 - \theta_{l-1,j}}{2}\right) \\ \overline{\theta_0} &= \widetilde{\theta}_0 - \varphi \end{aligned} \tag{19}$$

The fitness of each testing path T_j is then evaluated by using Eq. 17 . Since the probability $p(T_j)$ is a fitness function evaluating how much T_j is similar to our model of "normal" path, we exploited the *reject option* [19] to decide whether it fits or not, and, in this second case, the path is classified as "abnormal".

6. EXPERIMENTAL RESULTS

The presented system has been tested on different setups: in a public park in Reggio Emilia, Italy (Fig. 5) and in the Uni-

versity campus in Modena, Italy.



(a) (b)

Fig. 5. Examples of system working on actual cases with two cameras.

In this paper we will focus on the evaluation of higher-level modules, i.e. the consistent labeling and path classification modules.

The consistent labeling performance is highly correlated with the accuracy of the previous modules. For this reason, it has been tested on different video sequences acquired in real surveillance scenarios. Thus, tests have been performed considering several video sequences acquired during ordinary work days in different luminance conditions for two or three partially overlapped cameras. In acquiring the videos, no constraints have been imposed on people's trajectories or behaviors in order to test the system in actual conditions.

Three different situations have been considered for testing: the first refers to the case where a single person is moving from the field of view of a camera to another (*Single*), the second refers to a group of people entering in the field of view of one camera while they have been previously detected separated on an overlapped one (*Group Enter*) and finally the last case is when two or more people are entering one camera as separated while they have been already detected as a group in the other camera (*Group Inside*). Among the different videos analyzed, more than 1 hours of videos have been manually annotated to create ground truths. The achieved results are reported in the graph of Fig. 6. This graph highlights the performance of the system with respect to the different situations mentioned above (i.e., single person, group enter and group inside). Moreover, the last graph reports the overall accuracy of the system. The tests compared the accuracy with ground-plane homography only, with a Bayesian classifier with forward (or backward) contribution only (i.e., it does exploit neither priors nor backward (or forward) contribution), and with the complete HECOL approach. As it is evident, the complete approach has a higher overall accuracy.

The videos acquired and used for testing the consistent labeling have also been used to create the training set for the path analysis module. 88 paths sufficiently long (ranging from 100 to 400 frames, with an average framerate of 5 fps) have been extracted and used for training the classifier. The testing campaign has been carried out logging two hours of video at different time during an ordinary working day. We

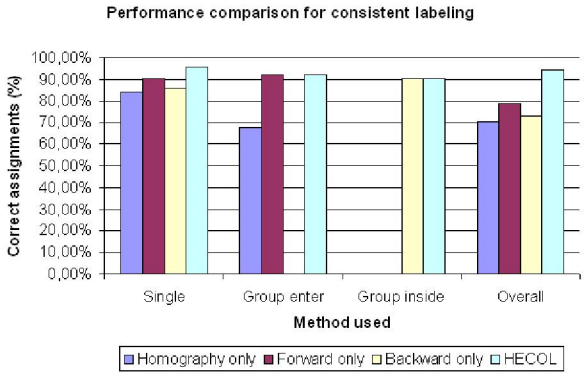
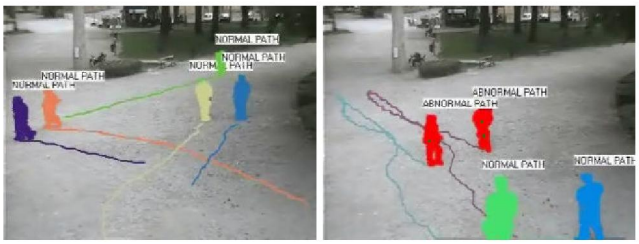


Fig. 6. Performance comparison.

collected two sets of trajectories: the first consisting of 121 trajectories with an average length of 90 points (corresponding to an average stay in the scene of 20 seconds) and the second of 135 trajectories with an average of 85 points (18 seconds). All the trajectories have been ground-truthed by several experts. The experts manually labeled the trajectories based on the training samples and their experience, and the final ground truth has been decided based on majority of votes. The experts divided the trajectories into 95 abnormal and 161 normal. The high number of abnormal trajectories is explained by the fact that the testing set has been balanced to test both types of classification. The classification rate is 100% for abnormal and 97.5% for normal. It is important to observe that the system correctly detects all the abnormal trajectories generating only false warnings in the case of normal behavior erroneously classified. Snapshots of the achieved results with superimposed the label assigned by the classifier are reported in Fi.g 7.



(a) All trajectories classified as normal. (b) Two trajectories (red objects) classified as abnormal.

Fig. 7. Examples of system classifying the observed trajectories

In conclusion, this paper describes the different modules composing our complete system for distributed video surveillance. At the time of this paper, the system includes a module for object detection and tracking from single fixed cameras, a module for consistent labeling between overlapped cameras, and a module for abnormal path detection exploiting the long paths created from the set of distributed cameras. Exper-

iments on each single module demonstrate the efficacy of the proposed approaches.

7. REFERENCES

[1] C. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Trans. on PAMI*, vol. 19, no. 7, pp. 780–785, July 1997.

[2] R.T. Collins, A.J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson, "A System for Video Surveillance and Monitoring: VSAM Final Report," *Technical report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University*, May 2000.

[3] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Trans. on PAMI*, vol. 22, no. 8, pp. 809–830, Aug. 2000.

[4] M. Isard and J. MacCormick, "Bramble: A bayesian multiple-blob tracker," in *Proc. Int. Conf. Computer Vision*, 2001, vol. 2, pp. 34–41.

[5] Oswald Lanz, "Approximate bayesian multibody tracking.," *IEEE Trans. on PAMI*, vol. 28, no. 9, pp. 1436–1449, 2006.

[6] J. Orwell, P. Remagnino, and G.A. Jones, "Multi-camera colour tracking," in *Proc. of Second IEEE Workshop on Visual Surveillance, (VS'99)*, June 1999, pp. 14–21.

[7] Z. Yue, S.K. Zhou, and R. Chellappa, "Robust two-camera tracking using homography," in *Proc. of IEEE Intl Conf. on Acoustics, Speech, and Signal Processing*, 2004, vol. 3, pp. 1–4.

[8] S. Khan and M. Shah, "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view," *IEEE Trans. on PAMI*, vol. 25, no. 10, pp. 1355–1360, Oct. 2003.

[9] Jinman Kang, I. Cohen, and G. Medioni, "Continuous tracking within and across camera streams," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition*, 2003, vol. 1, pp. I–267 – I–272.

[10] S.L. Dockstader and A.M. Tekalp, "Multiple camera tracking of interacting and occluded human motion," *Proc. of the IEEE*, vol. 89, no. 10, pp. 1441–1455, Oct. 2001.

[11] A. Mecocci and M. Pannozzo, "A completely autonomous system that learns anomalous movements in advanced videosurveillance applications," in *IEEE International Conference on Image Processing*, Sept 2005, vol. 2, pp. 586–589.

[12] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts and shadows in video streams," *IEEE Trans. on PAMI*, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.

[13] R. Cucchiara, C. Grana, G. Tardini, and R. Vezzani, "Probabilistic people tracking for occlusion handling," in *Proc. of International Conference on Pattern Recognition (ICPR 2004)*, Aug. 2004, vol. 1, pp. 132–135.

[14] S. Calderara, R. Melli, A. Prati, and R. Cucchiara, "Reliable background suppression for complex scenes," in *Proceedings of ACM Workshop on Video Surveillance and Sensor Networks (ACM VSSN 2006), Algorithm Competition*, 2006.

[15] A. Prati, I. Mikic, M.M. Trivedi, and R. Cucchiara, "Detecting moving shadows: Algorithms and evaluation," *IEEE Trans. on PAMI*, vol. 25, no. 7, pp. 918–923, July 2003.

[16] R. Cucchiara, C. Grana, A. Prati, and R. Vezzani, "Probabilistic posture classification for indoor surveillance," *IEEE Trans. on Systems, Man, and Cybernetics - Part A*, vol. 35, no. 1, pp. 42–54, Jan. 2005.

[17] C. Brauer-Burchardt and K. Voss, "Robust vanishing point determination in noisy images," in *Proc. of Int'l Conference on Pattern Recognition*, 2000, vol. 1, pp. 559–562.

[18] R.A. Fisher, "Dispersion on a sphere," *Proc. Roy. Soc. London Ser. A.*, vol. 217, pp. 295–305, 1953.

[19] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer–Verlag, 2006.

[20] Bhattacharyya A., "On a measure of divergence between two statistical populations defined by probability distributions," *Bulletins of Calcutta Math Society*, vol. 35, pp. 99–109, 1943.

[21] S. Calderara, R. Cucchiara, and A. Prati, "Detection of abnormal behaviors using a mixture of von mises distributions," in *in press in Proceedings of IEEE International Conference on Advanced Video and Signal based Surveillance (IEEE AVSS 2007)*, 2007.

[22] A.P. Reynolds, G. Richards, and V.J. Rayward-Smith, *The Application of K-Medoids and PAM to the Clustering of Rules*, vol. 3177/2004, pp. 173–178, Springer Berlin / Heidelberg.