Enabling technologies on hybrid camera networks for behavioral analysis of unattended indoor environments and their surroundings

Giovanni Gualdi, Andrea Prati, Rita Cucchiara
Dipartimento di Ingegneria dell'Informazione
University of Modena and Reggio Emilia
Italy
{giovanni.gualdi, andrea.prati, rita.cucchiara}@unimore.it

Edoardo Ardizzone, Marco La Cascia,
Liliana Lo Presti, Marco Morana
Dipartimento di Ingegneria Informatica
University of Palermo
Italy
{ardizzion, lacascia}@unipa.it, {lopresti,
morana}@dinfo.unipa.it

ABSTRACT

This paper presents a layered network architecture and the enabling technologies for accomplishing vision-based behavioral analysis of unattended environments. Specifically the vision network covers both the attended environment and its surroundings by means of hybrid cameras. The layer overlooking at the surroundings is laid outdoor and tracks people, monitoring entrance/exit points. It recovers the geometry of the site under surveillance and communicates people positions to a higher level layer. The layer monitoring the unattended environment undertakes similar goals, with the addition of maintaining a global mosaic of the observed scene for further understanding. Moreover, it merges information coming from sensors beyond the vision to deepen the understanding or increase the reliability of the system. The behavioral analysis is demanded to a third layer that merges the information received from the two other layers and infers knowledge about what happened, happens and will be likely happening in the environment. The paper also describes a case study that was implemented in the Engineering Campus of the University of Modena and Reggio Emilia, where our surveillance system has been deployed in a computer laboratory which was often unaccessible due to lack of attendance.

1. INTRODUCTION

Giving a quick glance to the panorama of information and communication technologies within research, development and production, it is quite clear the strong push on topics such as wireless communications, mobile computing, distributed networks, sensor networks and on top of them, high-level inferences on the huge amount of data produced with such technologies. Within such ferment, computer vision plays an ever increasing role, and as a specific branch of it, in recent times behavioral analysis has gained more and more attention thanks to two different enabling "technologies": on one side the rich variety of the sensed data (in the meaning of quality, quantity, multi-modality, distribution, etc.), on the other side the advances in machine learning and pattern recognition which can effectively process them. The research results are definitely promising and new and deeper level of behavior understanding are continuously uncovered; actually, it seems really hard to define a reasonable limit, if there is any, to the degree of behavior understanding which machines might attain.

If research is pushing forward, through this paper we claim that the technology is also mature enough to successfully step from research to development in the analysis of simple (or evident) behaviors which are characterized by well defined constraints and goals.

This paper specifically deals with the enabling technologies, mostly from computer vision and distributed sensor networks, that are used as building blocks of an automatic system for the attendance of indoor environments, which can successfully replace the tedious (and costly) human-based activity often limited to very low-level people monitoring (people positioning, counting, evident abuses of devices, thefts, etc). The field of applicability is wide: it is enough to quote the attendance of shops, libraries, labs, data-centers, etc. Many times such environments are left unattended, since the cost of a human attendance would be definitely bigger than the cost of the problems deriving from its lack.

The proposed architecture is based on a double-layer of camera networks: the first monitors the target environment, the second one its surroundings. Such outer layer does not simply extend the domain of observation, but increases the

range of achievable understanding and inference of the system. The employed cameras are of different typologies (fixed and PTZ - Pan-Tilt-Zoom), each assigned to complementary tasks. Moreover sensors beyond the vision could be used to deepen the understanding or increase the reliability of the system. The paper also describes a case study that very recently was implemented in the Engineering Campus of the University of Modena and Reggio Emilia, where our surveillance system has been introduced in a computer laboratory which was often unaccessible due to lack of attendance.

Several works on indoor [16, 42] or outdoor [20, 26, 10, 12, 19] surveillance are available in the literature. In our work, in order to obtain system scalability and efficient resource allocation, Wireless Multimedia Sensor Networks (WMSNs) are used. WMSNs [1] are enabling several new applications such as: multimedia surveillance, traffic avoidance and control systems, environmental monitoring, person locator services and many others. This type of network allows enhancing [14, 18] traditional monitoring and surveillance systems by using multi-sensor data to obtain multi-resolution and multi-source descriptions of the same scene. Moreover onboard processing algorithms allow to reduce bandwidth consumption. Hence, it is possible to obtain faster and more reliable systems but it is necessary to develop architectures to control distributed and collaborative information processing. Beside usual wired distributed systems [33, 43], there have been several studies on WMSNs. In [28] a model to decompose logical surveillance functionalities into a set of modules (e.g. tracking and classification of objects) is proposed. Each module is then optimally allocated among a set of physical processing units structured in a hierarchical surveillance network. Chu et al. [11] propose a system which uses onboard camera processing to filter out uninteresting events. A factor-graph-based resource allocation algorithm is then used to move cameras mantaining local and peripheral knowledge of new events.

This works does not claim to step into new surveillance fields or scenarios, but to propose and clarify a rendez-vous architecture and application for all these enabling technologies, which are mature to produce real solutions to real problems, fact that is confirmed as more and more world-wide technology companies and corporations are opening up to the video surveillance as part of their business areas.

The paper has the following structure: after a general architecture overview (section 2.1), a detailed description of each single layer will be given: perimeter vision layer (section 2.2), environment vision layer (section 2.3) and reasoning vision layer (section 2.4). We want to underline that what is presented in these sections is (and wants to be) a general and open description of an architecture, which can be specifically implemented in several different ways, depending on many factors and constraints (technological, economic, legal). Indeed, section 3 will give implementation details of the case study we deployed in our campus, providing the results obtained within such environment. Conclusions will follow.

2. SYSTEM ARCHITECTURE

2.1 An Overview

The general system architecture, depicted in Fig. 1, is divided into three main layers, namely: perimeter, environment and reasoning.

The *perimeter layer* deals with the surveillance of the surroundings of the environment to be attended. Usually (but not necessarily) it is based on nodes which are positioned outdoor. Each node is self standing and comprises a processing unit (typically an embedded platform), a camera, a wireless network interface and a power supply. The layer built upon such nodes forms a distributed sensor network that performs video surveillance tasks: object detection (with particular focus on people detection and stationary - possibly abandoned - objects), tracking (with consistent labeling [21, 9]) and people entrance/exit logging. This last task is particularly important, since this information will be handed off to the reasoning layer which will make inferences over the attended area and the people interacting around it. The architecture is designed in a way that the amount of information sent to the reasoning layer can be tuned according to wireless network capabilities and privacy restrictions: from a minimum level of just textual information about detected objects (e.g. trajectory points), to multi-media data (e.g. text, images, video). Since the nodes of this layer might be subject to frequent re-positioning due to environmental changes (very likely in outdoor setups) or privacy/law restrictions, it is requested that the architecture (hardware and software) on which they rely is easily and quickly deployable: particular attention is paid in a modular and scalable design and any initialization process (e.g. geometrical calibration) must be automatic or computer-assisted as much as possible. Moreover, the implemented algorithms must also take into account the fact that the processing power of the embedded processing units is limited compared to general processing units.

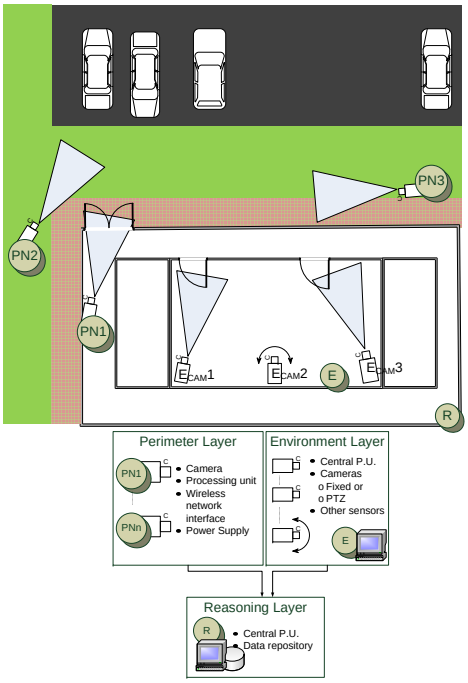

Figure 1: Scheme of the overall architecture

The *environment layer* deals with the surveillance of the environment area to be attended. Differently from the perimeter layer, is based on cameras which are not supposed to be repositioned. Therefore they are wire-connected to a central processing unit and the requirement of easy and quick

deployment is here loosened. In our implementation, the employed cameras are of two different kinds, which play complementary roles: fixed and PTZ cameras; but nothing hinders the architecture to exploit other kind of cameras (e.g. omni-directional) or even different sensors.

From a high level view, this layer extends the tasks that were requested in the perimeter layer. Beyond the people tracking, consistent labeling and entrance/exit monitoring, also more advanced tasks are performed; for instance, active tracking of moving objects through PTZ cameras: this is particularly helpful in case an object is leaving the field of view of the fixed cameras. Additionally, the PTZ cameras are used for a further task: a complete/partial mosaic of the environment [37] is built and kept updated (excluding the moving objects), and compared to the previous versions of it: this allows to detect anomalies in the environment, such as object misplacements or disappearances. The data produced by this layer is handed off to the reasoning layer.

The *reasoning layer*, which represent the core of the behavioral analysis, merges the information received from the two other layers to infer knowledge about what happened, happens and is likely to happen in the environment. Specifically, regarding the *present* (what happens), the reasoning layer provides an on-line people counter (restricted to the borders of the indoor environment), a people tracker and an analysis of the status of the environment infrastructure (missing / misplaced objects). Regarding the *past* (what happened), the layer offers logging about all the people that came in contact with the environment, offering trajectories, inferred information (e.g. interactions with other people or with infrastructures of the environment) and recorded visual data. Regarding the *future* (what will likely happen), the layer will infer, merging together the geometrical data and the perimeter observations, who is probably approaching or leaving the environment. Of course, these three pieces of information can be interrelated in order to deepen the knowledge over the environment (e.g. in case of misplaced object, it is possible to understand who interacted with it using trajectory analysis and, via visual data, understand what really happened).

2.2 The Perimeter Layer

The surveillance of the surroundings of the target environment is demanded to the perimeter vision layer, that is a distributed network made of a minimal set of wirelessly interconnected nodes (smart cameras). Each node consists of a micro-controller, a radio-communication device, a camera [41] and a power supply. In order to realize a self-standing device, it is even feasible to power up the nodes through solar panel [5].

In distributed video surveillance system, extraction of meaningful information from remote cameras to detect abnormal situations is a challenging task. Even worse, since data transmission is power consuming and the network bandwidth is limited, data processing is usually done locally and only aggregated data are sent over the network; moreover, smart cameras must process the acquired frames in real-time and independently from each other in order to guarantee system scalability. As shown in Fig. 2, one node is elected as master, with the task to aggregate data provided by processing nodes, to compute the consistent labeling, to handle the communication with the reasoning layer and to manage the insertion of new nodes in the layer.

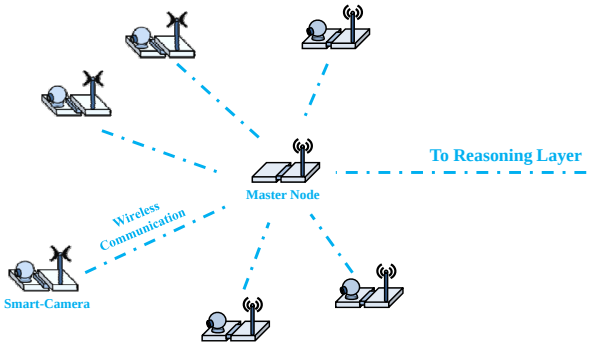

Figure 2: Scheme of the Perimeter Vision Layer

Tasks of each node can be distinguished between initialization and operational tasks. *Initialization tasks* regard basically two things: the background initialization, that will be used then for object segmentation through background suppression, and the geometry computation needed for the consistent labeling. There is a huge literature about background modeling but in this layer it is important to use a method that can successfully deal with typical outdoor challenges (e.g. illumination and environmental changes); under these conditions, the most suitable approach seems to be a statistical modeling [20, 40, 30]. Regarding the computation of the geometry needed for consistent labeling, [21, 9, 4], the approaches generally have a few geometrical requirements that need to be taken in consideration before the deployment of the system. For example [17] uses a light-weight solution that represents a good fit for limited computational power of the perimeter layer nodes; for a successful solution of the ambiguity over the tracks, this solution requires the contact of the observed object with the ground plane (Z=0) to be visible at least from one point of view. If this is satisfied, it is possible to locate the object position over the 2D plan of the environment through homography. There are several off-line methods to learn homography: similarly to what proposed in [9]: generally they require the collection of a certain amount of correspondences and methods to compute off-line a reliable homography. On the other hand, as aforementioned, the perimeter layer must be easy and fast to deploy, therefore an assisted geometry computation is advised; in [24], an on-line method to learn homography between pairs of camera views is presented. Using an on-line geometry learning is useful for the whole architecture in order to realize a scalable system that can dynamically change its deployment by adding, removing or repositioning smart cameras. Thus, this approach can be used not only for homography initial computation but also for its continuous refinement. It is important to state that parameters refinement ought to involve only few nodes at a time so that the rest of the system is still active for performing the regular surveillance tasks.

Operational tasks deal with object segmentation and detection. Each peripheral node acquires video tuning the resolution and the frame rate in order to cope with the limited resources of the micro-controller. The work in [35] provide a good review on the several background suppression techniques; the adopted solution is to use statistical methods to detect foreground but also midground [3], i.e. all those objects that become stationary in the scene. This allows to

rapidly highlight suspicious abandoned objects [38]. The object detection is forwarded to the master node that performs, by means of the site geometry, consistent labeling which can also solve partial/total occlusions of moving objects under single views [39].

The communication with Reasoning Layer is handled by the master node which initially provides, thanks to the geometry computation, an understanding of the position of the cameras and their fields of view. Every homography refinement is dispatched to the reasoning layer as well. Depending on the system set up, the perimeter vision layer can tune the amount and frequency of data regarding the object detection that is sent to the reasoning layer. The minimal information consists of the trajectory of an object, made of triples (object label, homography coordinate, time stamp) together with a visual descriptor (e.g. [36]) sent just once at the exit of the object from the scene. In fact, the descriptor introduced in [36] incrementally accumulates visual information about the object; therefore, when the object leaves the scene the descriptor contains information on its whole "history" in the scene. If the system can support richer data exchange (from a technological and/or lawful point of view), the minimal information could be updated with higher frequency and enriched with visual data (images, video clips, etc.).

2.3 The Environment Layer

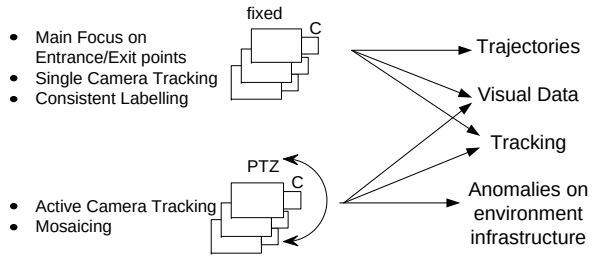

Figure 3: Scheme of the environment layer

The *fixed cameras* are positioned inside the environment so that, primarily, all the entrance/exit points are kept under observation and, secondarily, the widest possible area is covered by their fields of view. Since the cameras are fixed, the object segmentation and the single-view tracking can be exploited using the background suppression: differently from the perimeter layer, the vision in this layer is indoor therefore simpler statistical approaches like [15] could be equally effective but more efficient; to deal with object occlusions within the same view, some appearance models (based on color, texture, contour, etc) can be exploited [17, 23]. Vision-based people counting at gates has been widely explored in the literature [29, 2, 6], but, apart from ad-hoc approaches, it could be interpreted also as the outcome of a correct people tracker (see [31] for a survey) which observes all the entrance/exit gates of an environment, that is our case. To increase the accuracy of the counting, additional sensors could be deployed, as will be detailed in our case study in section 3. Regarding the consistent labeling the same consideration made for the perimeter layer are valid in this layer, with the only difference that in this layer, being the camera set-up more stable, an on-line homography learning is not really necessary, and a more precise off-line procedure can be employed.

The *PTZ cameras* have a double task: primarily the PTZ capability of spanning over a wide view of the environment is used to build up and then keep updated a global mosaic of the observed scene (or a portion of it) [37]. The bottom line idea is to compare the actual mosaic of the environment against a *background mosaic image*, in order to highlight differences that might point out changes on the infrastructure, detecting ordinary objects moves (e.g. chair moved), or extra-ordinary ones (e.g. closet door left opened, device missing, etc). Once the mosaic extraction can be referred against absolutes coordinates (or PTZ coordinates), the mosaic background calculation does not differ much from a background image calculation in single view video and can be updated using statistical and/or selective processes. On the top of this, since the environment layer can also exploit the object tracking information from the fixed cameras, the mosaic update could be calculated just in the portion of the views where it is known to have no persons/objects. In case the mosaic differencing detects an infrastructure anomaly, it is possible to run a scale-rotation invariant and occlusion-robust object recognition ([25] for example) in order to detect if the highlighted object simply moved within the scene or disappeared from it.

While the mosaicing mode is a default action of the PTZ cameras, the secondary PTZ task is triggered only when a tracked object is spatially close to the borders of the field of view of the fixed cameras (let consider that through homography and geometric calibration it is possible to recover accurately the field of view of each camera). In this case the closest PTZ camera (depending on the adopted PTZ management policy, it could be more than one), is commanded to target the area where the border-line object is: again, the knowledge of the object position relative to the homography, through simple coordinate transformations, can provide the information for PTZ guidance [7, 44]. Once the PTZ is approximately pointing to the direction where the object is, an active camera tracking is started. In literature the active camera tracking is usually faced using motion compensation, affine transformations or depth sensors [32, 22, 27]. We also propose an uncalibrated approach based on a kernel-based tracking (similar to [8, 13]), or a particle filter tracking [34] that feeds a PTZ guidance module in order to keep the tracked person within the field of view of the PTZ camera (the so-called *person following* task [37]). These approaches need an object model that might be initially provided with the object visual features extracted from the fixed camera tracking, and then possibly refined during the active camera tracking. The PTZ active camera tracking is suspended when the tracked object enters again the field of view covered by the fixed cameras.

The environment layer, similarly to the perimeter one, forwards the homography, with camera positions and fields of view to the reasoning layer. Being this layer wirely connected to the network, the data regarding moving objects is forwarded by default with the maximum frequency (i.e. equal to the frame rate), and maximum degree of detail (trajectories, object descriptor, images, clips, etc). As will be better detailed in section 2.4, whenever a new object appears, its label is not assigned merely on the information provided by the environment layer, but the final assignment is decided by the reasoning layer, with the goal to keep a consistent labeling with the tracking performed in the perimeter

layer.

2.4 The Reasoning Layer

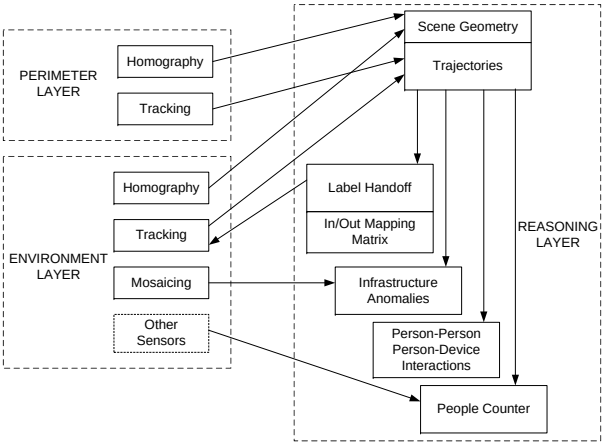

Figure 4: Scheme of the reasoning layer and the interactions with the other layers

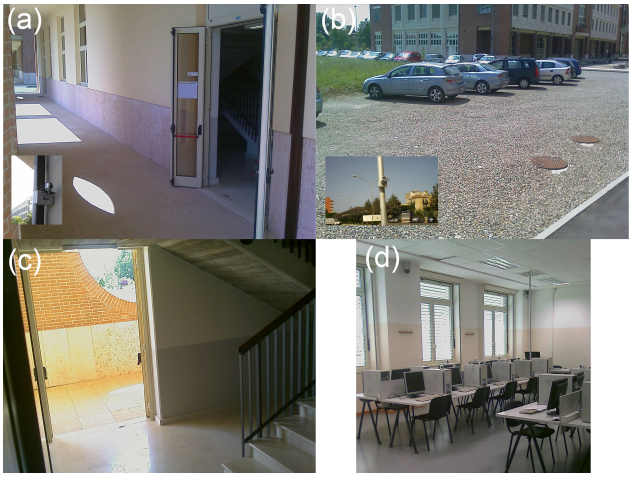

Figure 5: (a,b,c) are taken from 3 of the 4 perimeter nodes. (d) shows the lab and two of the three cameras.

The reasoning layer, represented in its functionalities in Fig. 4, infers behaviors in the attended environment by using data provided by perimeter and environment layers. As it can be seen from this scheme, the central entity is represented by the trajectories of the detected objects, which are basically lists of triples (object label; coordinate; time stamp) plus a descriptor based on the object features (typically color, texture, contours). The knowledge about the geometry of the cameras is necessary in order to correctly interpret the coordinates received by the vision layers.

The *Label Handoff* module has the task to correlate objects detected in the perimeter with the ones appeared in the monitored environment. In order to do this, the module makes use of three matrices (whose values are manually or automatically learnt) which correlate all the possible exit points of the perimeter layer (rows of the matrix) with all the

Figure 6: Mosaic built up after an horizontal span of the PTZ inside the lab

possible entrance points (columns of the matrix); the matrices are $IOTM$ (in/out transition matrix), $IOMM$ (in/out mean (time) matrix) and $IOVM$ (in/out variance (time) matrix); each element $(i; j)$ is explained as follows:

- $IOTM(i, j)$: binary value depending whether it is possible to reach entrance j from exit i
- $IOMM(i, j)$: mean time μ that it is taken to reach entrance j from exit i, modeling the trip time as a Gaussian
- $IOVM(i, j)$: variance σ^2 of the Gaussian modeling the trip time

Defining a generic new track $\widetilde{\tau}^e$ detected by the environment layer (where the superscript indicates the involved layer - e or p for environment or perimeter, respectively), t_{in} and t_{out} as its entrance and exit times, P_{in} and P_{out} as the entrance and exit point identifiers and D as the descriptor of the track; defining $\Omega\left(D\left(\tau_a\right), D\left(\tau_b\right)\right)$ as descriptor similarity function between track τ_a and track τ_b, the label $L\left(\widetilde{\tau}^e\right)$ is determined as in equation 1, where $\mathcal{N}$ represents a Gaussian distribution.

To have a more robust match, the label assignment reported in equation 1 is performed only if the best match and the ratio between the best match and the second best match respectively exceed two thresholds (similarly to what is proposed by [25] for robust key point matching).

The module for *infrastructure anomalies* correlates possible missing/misplaced objects detected by the environment layer with the trajectories that are correlated to them by space-time proximity. The *interactions* module works in complete analogy, providing person-to-person or person-to-device interactions that are established again on space-time proximity.

The *people counter* module is a simple counting operation on the trajectories, and as aforementioned, can be made more robust through the use of additional sensing information.

3. CASE STUDY: LAB ATTENDANCE

The study proposed in this paper ended in the deployment of a real prototype used for the monitoring of a laboratory of the University of Modena and Reggio Emilia that is not utilized up to the present moment because of the lack of university personnel attendance monitoring the environment. This situation motivated our research work in order to develop a surveillance system for behavior analysis of the laboratory and its surroundings.

The perimeter layer deploys 4 nodes, 3 are placed outdoor, the other is indoor pointing to the main entrance gate of the building. Each node is based on a Single Board Computer

$$L\left(\widetilde{\tau}^{e}\right)=L\left(\overline{\tau}^{p}\right)\quad|\quad\overline{\tau}^{p}=\underset{\tau_{k}^{p}}{\arg\max}\;\;\delta\cdot\Omega\left(D\left(\tau_{k}^{p}\right),D\left(\widetilde{\tau}^{e}\right)\right)\cdot\mathcal{N}\left(t_{in}\left(\widetilde{\tau}^{e}\right)-t_{out}\left(\tau_{k}^{p}\right)|\mu,\sigma\right)\tag{1}$$

$$\delta=IOTM\left(P_{out}\left(\tau_{k}^{p}\right),P_{in}\left(\widetilde{\tau}^{e}\right)\right)\qquad\mu=IOMM\left(P_{out}\left(\tau_{k}^{p}\right),P_{in}\left(\widetilde{\tau}^{e}\right)\right)\qquad\sigma^{2}=IOVM\left(P_{out}\left(\tau_{k}^{p}\right),P_{in}\left(\widetilde{\tau}^{e}\right)\right)\tag{2}$$

called Stargate (produced by CrossBow Technology) based on a xScale 400Mhz (PXA255) Intel CPU. It supports standard interfaces like USB and PCMCIA, used respectively for video input through web-cameras and Wi-Fi interfaces. Tests were successfully performed also with a GPRS radio-mobile modem.

Figure 7: CrossBow Stargate

The video acquisition is performed at QVGA or QQVGA resolution and 10fps. Foreground and midground objects are detected by using the method presented in [38]. On top of object segmentation, smart cameras perform local (single-view) tracking, with object split/merge handling as described in [39]. Nodes estimate the trajectory of each object on the image plane by approximating it to a piecewise linear function. Vertices of this function correspond to the points in which the object changes the direction. The local track data is forwarded to a master node which performs consistent labeling. The homography is computed at set-up time and then refined on-line as described in [24] while the appearance model is described in [36].

The environment layer deploys 2 fixed cameras each observing one of the two entrance doors of the lab, and a PTZ camera, similarly to what was depicted in fig. 1. The fixed cameras perform object segmentation and single-view tracking according to [15]. After a manual off-line calibration, consistent labeling is performed according to [9]. The PTZ-based mosaic is build with an optical-flow compensation, based on [37], and the mosaic differencing is performed with the same background modeling used over the fixed cameras. The mosaic is performed on a 180° wide pan angle, and 90° wide tilt angle (from horizontal to down-ward vertical orientation). The mosaicing module is also provided with a binary mask which highlights the presence of devices (printer, projector, white board, PCs): mosaic differencing that detects objects in correspondence of the objects of such binary mast will trigger an infrastructure anomaly. So far the anomalies do not make distinction between missing and misplaced objects, as we are working in the implementation of the SIFT-based object recognition part, for detecting object misplacements. The active camera tracking is based on particle filter tracking [34], moving the camera only in the pan-tilt directions (no zooming) in order to keep the tracked object in the center of the PTZ image plane.

The reasoning layer is provided with manually initialized In/Out matrices of size 4x2. The communication is all TCP/IP based and the trajectories data are stored on Microsoft SQL server. The people counting is performed based on vision-based people counting, and checked in its correctness through a PIR based people counter as described in [18]. The feature descriptor used throughout the three modules (for label handoff and for PTZ active tracking initialization) is based on [36].

The performance evaluation of each single component can be found in the above-mentioned papers, while the evaluation of the overall system over the three layers is still under progress, since not enough data has been collected yet.

4. CONCLUSIONS

The paper presents a panoramic view over the enabling technologies that, through an ad-hoc architecture, are already mature for being deployed in successful applications for behavior analysis of unattended indoor environments. The presented architecture is based on a scalable and modular approach and purposely separates vision from reasoning and environment sensing from surroundings sensing. We believe that in the next future many application similar to what is described here, with modifications and additions, will seriously step into the market of surveillance solutions and services. The deployment of a real prototype at the University of Modena and Reggio Emilia allowed to design and progressively refine the system architecture using a real-word scenario. In our previous works we have evaluated the performance of each single component independently from each other, while the evaluation of the overall three-layer system is still in progress. Our next future commitment is to collect data for experimental results, both on robustness of the behavioral analysis and on the effectiveness of the architecture deployment.

5. ACKNOWLEDGMENTS

This work is supported by the project FREE SURF (FREE SUrveillance in a pRivacy respectFul way), funded by MIUR (project nr. 2006099482).

6. REFERENCES

[1] I. Akyildiz, T. Melodia, and K. Chowdury. Wireless multimedia sensor networks: A survey. *Wireless Communications, IEEE [see also IEEE Personal Communications]*, 14(6):32–39, December 2007.

[2] A. Albiol, I. Mora, and V. Naranjo. Real-time high density people counter using morphological tools. *Intelligent Transportation Systems, IEEE Transactions on*, 2(4):204–218, Dec 2001.

[3] S. Apewokin, B. Valentine, L. Wills, S. Wills, and A. Gentile. Midground object detection in real world video scenes. *IEEE Conference on Advanced Video and Signal based Surveillance AVSS '07.*, 2007.

[4] M. Balcells, D. DeMenthon, and D. Doermann. An appearance-based approach for consistent labeling of humans and objects in video. *Pattern Anal. Appl.*, 7(4):373–385, 2004.

[5] N. Banerjee, M. Corner, and B. Levine. An energy-efficient architecture for dtn throwboxes. *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, pages 776–784, May 2007.

[6] D. Biliotti, G. Antonini, and J. P. Thiran. Multi-layer hierarchical clustering of pedestrian trajectories for automatic counting of people in video sequences. In *WACV-MOTION '05: Proceedings of the IEEE Workshop on Motion and Video Computing (WACV/MOTION'05) - Volume 2*, pages 50–57, 2005.

[7] A. D. Bimbo and F. Pernici. Uncalibrated 3d human tracking with a ptz-camera viewing a plane. In *Proc. 3DTV International Conference: Capture, Transmission and Display of 3D Video (3DTV-CON 08)*, 2008.

[8] G. Bradski. Real time face and object tracking as a component of a perceptual user interface. In *Proc. of WACV*, pages 214 – 219, 1998.

[9] S. Calderara, R. Cucchiara, and A. Prati. Bayesian-competitive consistent labeling for people surveillance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):354–360, 2008.

[10] T. P. Chen, H. Haussecker, A. Bovyrin, R. Belenov, K. Rodyushkin, A. Kuranov, and V. Eruhimov. Computer vision workload analysis: Case study of video surveillance systems. *Intel Technology Journal - Compute-intensive, highly parallel applications and uses*, 9:109 – 118, 2005.

[11] M. Chu, J. Reich, and F. Zhao. Distributed attention in large scale video sensor networks. *Intelligent Distributed Surveilliance Systems, IEE*, pages 61–65, Feb. 2004.

[12] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89:1456 – 1477, 2001.

[13] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 2:142–149 vol.2, 2000.

[14] R. Cucchiara. Multimedia surveillance systems. In *VSSN '05: Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, pages 3–10, New York, NY, USA, 2005. ACM.

[15] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, Oct. 2003.

[16] R. Cucchiara, C. Grana, A. Prati, and R. Vezzani. Computer vision system for in-house video surveillance. *Vision, Image and Signal Processing, IEE Proceedings* -, 152(2):242–249, April 2005.

[17] R. Cucchiara, C. Grana, G. Tardini, and R. Vezzani. Probabilistic people tracking for occlusion handling. In *Proceedings of IAPR International Conference on Pattern Recognition (ICPR 2004)*, pages 132–135, Aug. 2004.

[18] R. Cucchiara, A. Prati, R. Vezzani, L. Benini, E. Farella, and P. Zappi. Using a wireless sensor network to enhance video surveillance. *Journal of Ubiquitous Computing and Intelligence (JUCI)*, 1:1–11, 2006.

[19] G. Foresti. A real-time system for video surveillance of unattended outdoor environments. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(6):697–704, Oct 1998.

[20] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:809 – 830, 2000.

[21] S. Khan and M. Shah. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(10):1355–1360, Oct. 2003.

[22] K. K. Kim, S. H. Cho, H. J. Kim, and J. Y. Lee. Detecting and tracking moving object using an active camera. *Advanced Communication Technology, 2005, ICACT 2005. The 7th International Conference on*, 2:817–820, 2005.

[23] X. Li. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26:1531–1536, 2004. Member-Alper Yilmaz and Fellow-Mubarak Shah.

[24] L. lo Presti and M. L. Cascia. Real-time estimation of geometrical transformation between views in distributed smart-cameras systems. *International Conference on Distributed Smart Cameras - ICDSC08*, in press.

[25] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[26] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson. Wireless sensor networks for habitat monitoring. In *WSNA '02: Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, pages 88–97. ACM, 2002.

[27] A. Maki, P. Nordlund, and J.-O. Eklundh. Attentional scene segmentation: integrating depth and motion. *Comput. Vis. Image Underst.*, 78(3):351–373, 2000.

[28] L. Marcenaro, F. Oberti, G. Foresti, and C. Regazzoni. Distributed architectures and logical-task decomposition in multimedia surveillance systems. *Proceedings of the IEEE*, 89(10):1419–1440, Oct 2001.

[29] O. Masoud and N. Papanikolopoulos. A novel method for tracking and counting pedestrians in real-time using a single camera. *Vehicular Technology, IEEE Transactions on*, 50(5):1267–1278, Sep 2001.

[30] A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2:II–302–II–309 Vol.2, June-2 July 2004.

[31] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104(2):90–126, 2006.

[32] D. Murray and A. Basu. Motion tracking with an

active camera. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(5):449–459, May 1994.

[33] D. Ostheimer, S. Lemay, M. Ghazal, D. Mayisela, A. Amer, and P. F. Dagba. A modular distributed video surveillance system over ip. In *CCECE*, pages 518–521. IEEE, 2006.

[34] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I*, pages 661–675, London, UK, 2002.

[35] M. Piccardi. Background subtraction techniques: a review. *IEEE International Conference on Systems, Man and Cybernetics*, 4:3099 – 3104, 2004.

[36] M. Piccardi and E. D. Cheng. Multi-frame moving object track matching based on an incremental major color spectrum histogram matching algorithm. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, 2005.

[37] A. Prati, F. Seghedoni, and R. Cucchiara. Fast dynamic mosaicing and person following. In *Proc. of International Conference on Pattern Recognition (ICPR 2006)*, volume 4, pages 920–923, Aug. 2006.

[38] L. L. Presti and M. L. Cascia. Real-time object detection in embedded video surveillance systems. *9th International Workshop on Image Analysis for Multimedia Interactive Services, IEEE Proceedings*, 2008.

[39] L. Snidaro, C. Micheloni, and C. Chiavedale. Video security for ambient intelligence. *Systems, Man and Cybernetics, Part A, IEEE Transactions on*, 35:133 – 144, Jan. 2005.

[40] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. *Conference on Computer Vision and Pattern Recognition '99,IEEE Computer Society*, 2:246 – 252, 1999.

[41] M. Valera and S. A. Velastin. Intelligent distributed surveillance systems: a review. *Vision, Image and Signal Processing, IEEE Proceedings*, 152:192 – 204, 2005.

[42] T. Yang, F. Chen, D. Kimber, and J. Vaughan. Robust people detection and tracking in a multi-camera indoor visual surveillance system. *Multimedia and Expo, 2007 IEEE International Conference on*, pages 675–678, July 2007.

[43] X. Yuan, Z. Sun, Y. Varol, and G. Bebis. A distributed visual surveillance system. In *AVSS '03: Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, page 199, Washington, DC, USA, 2003. IEEE Computer Society.

[44] X. Zhou, R. T. Collins, T. Kanade, and P. Metes. A master-slave system to acquire biometric imagery of humans at distance. In *IWVS '03: First ACM SIGMM international workshop on Video surveillance*, pages 113–120, 2003.