User detection through multi-sensor fusion in an AmI scenario

A. De Paola, M. La Cascia, G. Lo Re, M. Morana, M. Ortolani

In Proceedings of the 15th International Conference on Information Fusion, 2012, pp. 2502-2509

User Detection through Multi-Sensor Fusion in an AmI Scenario

Alessandra De Paola, Marco La Cascia, Giuseppe Lo Re, Marco Morana and Marco Ortolani

University of Palermo
Viale delle Scienze, ed. 6 - 90128 Palermo, Italy
E-mail: {alessandra.depaola, marco.lacascia, giuseppe.lore, marco.morana, marco.ortolani}@unipa.it

***Abstract*—Recent advances in technology, with regard to sensing and transmission devices, have made it possible to obtain continuous and precise monitoring of a wide range of qualitatively diverse environments. This has boosted the research on the novel field of Ambient Intelligence, which aims at exploiting the information about the environment state in order to adapt it to the user's preference. In this paper, we analyze the issue of detecting the user's presence in a given region of the monitored area, which is crucial in order to trigger subsequent actions. In particular, we present a comprehensive framework that turns data perceived by sensors of different nature, and with possible imprecision, into higher-level information; a case study derived from an actual implementation of the system regarding the management of an office environment is also described, and experimental results are presented.**

I. Introduction

Ambient Intelligence (AmI) is a new paradigm in Artificial Intelligence that introduces a shift in perspective as regards the role of the end user [1]. Unlike other well established approaches, such as the human-in-the-loop design, where the contribution resulting from the exploitation of the human factor is limited to facilitate the system design process, or to infer more accurate models for the environment state, Ambient Intelligence aims to fully integrate the user's preference into the system. In this respect, the basic intrinsic requirement of any AmI system is the presence of pervasive sensory devices [2], which are essential to ensure context-aware reasoning in order to act upon the environment, modify its state, and react to user-driven stimuli. Today's advances in technology allow for cheap and unintrusive sensors that may be profitably employed as a distributed sensory means permeating the whole environment under observation. In particular, Wireless Sensor Networks (WSNs) [3] allow to get precise and continuous monitoring of the physical quantities of interest; not only does this novel technology allow to perform remote sensing without causing disruption, but it may also perform basic *in-network* preprocessing of sensed data thanks to the limited computational capabilities of the nodes.

Even if WSNs easily allow for low-level ambient sensing, basic nodes are not sufficient to perceive high-level features such as *who* is in the office or *what* this person is doing there (e.g. reading, talking, using their workstations, and so on); for this purpose, high-level vision sensors are needed. The growing attention on embedded vision-based techniques can be mainly attributed to the increasing availability of small devices capable of sensing the environment, performing onboard processing on captured data and exchanging it with other devices in a collaborative way. Some face detection (e.g., Viola-Jones face detector [4]) and face recognition (e.g., local binary patterns [5]) techniques have reached a good level of maturity, so we focused on their implementation on embedded systems, taking into account both hardware and software constraints.

Our work focuses on *sensing* the presence of the user by producing a description of the observed scene; in order to ensure system scalability and efficient resource allocation, a variant of WSNs is used, namely Wireless Multimedia Sensor Networks (WMSNs), which are characterized by the addition of video sensors. Face processing is performed on each node and extracted data is sent to a server which will make inferences over the people interacting around each observed area. However, in such a dynamic scenario, sensory data is likely to be biased by environmental noise and by the unavoidably imperfect nature of sensor devices, so it is convenient to adopt an approach able to cope with uncertainty for developing reasoning components. Our architecture is thus designed to make use of a probabilistic approach that allows to meet this requirement, and also to manage information fusion in a dynamic scenario. In particular, this work describes an approach based on Bayesian Networks for merging data coming from heterogeneous sensors, in order to improve the detection of the users' presence.

The remote, distributed sensory device thus acts as the termination of a centralized sentient reasoner, where actual intelligent processing occurs; sensed data is processed in order to extract higher-level information, carrying on symbolic reasoning on the inferred concepts, and producing the necessary actions to adapt the environment to the users' requirements. A set of actuators finally takes care of putting the planned modifications to the environment state into practice.

The paper is organized as follows: an analysis of related work will be given in Sect. II, and the proposed system will be fully described in Sect. III. A case study implemented at the Department of Computer Engineering of the University of Palermo will be shown and discussed in Sect. IV. Conclusions will follow in Sect. V.

II. Related Work

Wireless Multimedia Sensor Networks (WMSNs), i.e., *"networks of wirelessly interconnected devices that allow retrieving video and audio streams, still images, and scalar sensor data"* [6], are enabling several new applications such as multimedia surveillance sensor networks, environmental monitoring and many others. Traditional wired distributed monitoring systems [7], [8] are deeply constrained at design time so that a reengineering process often requires a great effort making the system static and unmodifiable. WMSNs extend traditional systems by using multiple sensors to perceive the environment from different, not necessarily predefined, viewpoints. Each sensor device is usually independent of other nodes and connected to the wireless network, therefore it can be moved, added or removed from the system without difficulty. In [9] the design and implementation of a distributed search system over a camera sensor network is described. Each node is an iMote2 sensor device that senses, stores and searches information. A network of dual-camera nodes is proposed by [10] to facilitate retrieval of misplaced objects in a home environment. Each platform consists of a low-power (MICAz mote equipped with CyclopsCam) and a high-power (iMote2 equipped with EnalabCam) camera sensor nodes. Some works on embedded solutions for face detection or recognition have also been proposed: a study for an embedded implementation of boosting-based face detection is described in [11]. In [12] the authors present an architecture to perform real-time face recognition using smart cameras. However many works are usually limited to the implementation of a single face processing task.

The face detection module we developed is based on Viola-Jones [4] face detector (VJFD), that is the most stable and used face detector both in academic and commercial systems. Face recognition step is based on Local Binary Pattern (LBP), an operator invariant to monotonic gray level and computationally efficient. In [13] a comparison of LBP, Eigenfaces and Fisherfaces descriptors is reported. To reduce the effects of different illumination conditions, we used the method presented in [14].

In WMSNs literature a lot of works are presented about sensor fusion algorithms, in order to cope with natural uncertainity related to this kind of technology. In [15] a distributed Bayesian algorithm is proposed to perform the sensor fusion task, and to solve the fault-event disambiguation problem in sensor networks. The authors of [16] propose a dynamic infrastructure based on Bayesian networks, for unifying information coming from a wireless network.

The authors of [17] propose an AmI system able to analyze simultaneous data coming from a heterogeneous network (virtual sensors, CCD cameras, probes, etc.), combining them in a unique and symbolic representation of what happens in the environment. In our work the fusion of information coming from different sensors, including RFId readers, ambient and video sensors is made through a Bayesian Network in which each piece of information is treated as a sensor signal regardless of the subsystem that has generated it.

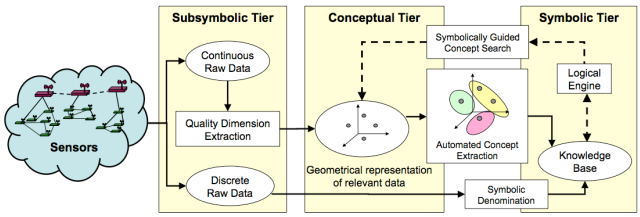

Fig. 1. The three-tier structure of a low level module.

III. System Overview

The architecture proposed in this work is inspired by the human nervous system, in which signals gathered by the peripheral system are filtered, aggregated and then sent to the central system for high-level processing. We consider as case study a home automation application instantiated for a work environment, with the aim to provide constant monitoring of the environmental conditions in the rooms of the teaching staff of our department. After presenting an overview of our multi-tier approach to knowledge representation, we go on to describe the designed WMSN, representing the peripheral system that permeates the environment, and allows for distributed data pre-processing; finally, this section outlines the modular structure of the intelligent system, where sensor fusion is performed.

A. Multi-tier Knowledge Representation

The proposed system is based on a multi-tier paradigm for performing knowledge extraction starting from sensory data [18]. As shown in Figure 1, this paradigm provides three tiers of knowledge representation, corresponding to different abstraction degrees. Starting from the rightmost block in the figure, knowledge is represented at *linguistic* level, where information is described symbolically via a high-level language, whose input is provided by a *conceptual* level where grounding of symbols occurs, and used to connect the system to the lower, *subsymbolic* tier, where sensory data is first acquired.

The subsymbolic tier processes the measurements collected by the pervasive sensory subsystem. As already mentioned, the purpose of the WMSN-based infrastructure is not limited to the basic gathering of sensed data, but comprises also a preliminary processing aimed at the selection of the relevant information. Sensed measurements can be classified into two main categories, namely continuous or discrete; data belonging to the former class is fed to the intermediate conceptual tier, where they will be provided with a representation in terms of continuous quality dimensions. On the other hand, discrete data is outright handed over to the symbolic tier, where a linguistic representation will be given.

At the conceptual tier, data is endowed with a geometrical representation that allows for a straightforward management of the notion of concept similarity, as long as a proper metric is chosen for the quality dimensions. Points populating the conceptual space, originally generated by the underlying measurement space, are represented as vectors, whose compo-

nents are the quality measurements of interest. Concepts thus naturally arise from the geometric space as regions, identifiable through an automated classification process, and points will belong to one of those regions. In our implementation the identification of regions associated to concepts occurs after a supervised training of the classifier. As will be detailed in the following, the classifier is also able to dynamically adjust its internal representation of the concepts based on direct and indirect feedback from the user.

The symbolic tier in each module produces a concise description of the environment by means of a high-level logical language. At this level, regions individuated inside the conceptual space are associated to a linguistic construct, thus identifying basic concepts, while relations necessary to infer more complex concepts are described through an opportune ontology. The gap between a concept and its linguistic description is filled through two separate mechanisms inspired to the work of [19]: an "automated concept extractor" deals with the translation of the regions in the conceptual space into symbolic elements, whereas a "symbolically guided concept search" identifies further points in the conceptual space as a consequence of the activation of some of the logical rules contained at the symbolic tier.

The created knowledge base is used to iterate the same knowledge extraction mechanisms at a higher abstraction level. In the considered case study, the concepts asserted at the symbolic tier are also employed for the activation of the control rules of the actuators, represented by the controllers of the heat, air conditioning, and lighting systems. Moreover, a subset of those rules is devoted to providing feedback to the WMSN in order to guide its self-maintenance activity; for instance, under steady environmental conditions, the higher tier will opt for a reduction of the sensor sampling rate in order to reduce the overall energy consumption.

B. Eyes on the User - The Sensory Component

WMSNs represent the sensory component of our system, that permeate the environment and allow for distributed data pre-processing. We regard the aggregation and selection of environmental data as analogous to the processing of perceptual signals occurring in the human nervous system. Some components of the peripheral system filter perceptual information by means of distributed processing among several neurons. A remarkable example is the processing of visual information occurring in the retina [20]: in the human eye, photoreceptors convert light into electrical signals that are passed to a network of retinal neurons, and are modified before being transmitted to gangliar neurons; eventually, they are handed to the optic nerve that carries the information up to the brain. The retinal neuron network does not restrict itself to carrying signals from photoreceptors, but rather combines them to obtain an aggregate heavily dependent on the spatial and temporal features of the original light signal.

In our architecture the terminal sensory component performing is represented by WMSNs pervasively deployed in the environment.

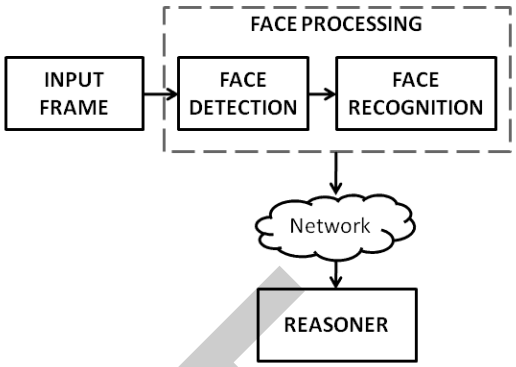

Fig. 2. Scheme of the face processing module.

In order to detect user presence, the most functional sensors, that is sensors that produce signals the most correlated with the signal representation of user's presence, are video sensors. From an exclusively functional point of view, a part from the particular deployment and hardware implementation, video sensors has to perceive high-level features such as *who* is in the office or *what* this person is doing. In particular, in this work, we describe the use of video sensor to detect user presence through a face recognition process.

Face processing is performed in two steps: firstly face detection is performed on the acquired frame, then a face is sent to the face recognition module obtaining the face id and the probability with which the id is assigned.

The framework proposed by Viola and Jones has been chosen since it represents the state of the art approach to face detection. Images are classified by evaluating the values of three simple *rectangular features*.

Each feature is scaled and shifted across all possible combinations (e.g., in a window of 24×24 pixel, 160,000 possible features are to be computed), however the use of an image representation called *integral image* allows the features to be computed very quickly in just a few references.

A variant of AdaBoost [21] is then used both to select the best features from the huge feature space (e.g., 160,000 rectangle features associated with each image sub-window) and to combine them to train the classifier. Computation time is further reduced by arranging the classifiers in a cascade, a decision tree, where a classifier at stage t is trained only on those examples which pass through all the previous stages. Thus, early stages of the cascade allow background regions of the image to be quickly discarded while spending more computation on promising regions.

Once a face has been detected and normalized for scale (i.e., 110×110 pixels) it is possible to proceed to the face recognition step.

Local Binary Pattern (LBP) [22] is a non-parametric kernel which summarizes the local spacial structure of an image. At a given pixel position, LBP is defined as an ordered set of binary comparisons of pixel intensities between the centre pixel and its eight surrounding pixels.

The LBP operator performs the labeling of image pixels by thresholding the $M \times N$ neighborhood of each pixel with the center pixel value $p(x, y)$. Pixel labels are obtained by considering the result of the thresholding as a binary number (i.e., 1 if $p(m, n) > p(x, y)$, 0 otherwise). In [23], LBP was used for face recognition. The histograms of the extracted labels are concatenated to form a global descriptor of a generic image. In order to preserve spatial relations between facial features, the authors define *spatially enhanced histograms* for encoding both the appearance and the spatial relations of facial regions.

Face recognition is performed by comparing LBP feature vectors. As discussed in [13], the probability p of a match given the distance can me obtained by using a gamma density functions to model the histograms of the distances of each couple of LBP-based face descriptors in a training set in case of match and mismatch.

C. Modular Architecture for Sensor Fusion

The proposed system is organized according to a hierarchical structure whose modules are combined together in order to carry on specific reasoning on the environment at different levels of abstraction and on different kinds of perceptions. The overall behavior mimics that of the human brain, where the emerging complex behavior is the result of the interaction among smaller subsystems. From the design point of view, the modular organization allows for the realization of a scalable sofware architecture, able to effectively manage the huge amount of sensory data.

In our modular architecture, the outcome of lower-level reasoning is fed into the upper levels, that deal with the integration of information originated by multiple lower-level modules. Each module independently measures environmental quantities and conceptualizes them.

Considering a particular scenario, the human language comprehension model, described in [20], provides a significant example of interaction patterns among specific areas of the brain.Different anatomic structures are devoted to different phases of language processing, the primary auditory cortex initially processes the auditory signals while at the same time the primary visual cortex processes the visual signals. In our architecture, an analogous example may be recognized in the modules devoted to detect user's presence: low-level modules independently reason about noise level (auditory input) and a visual recognition of user's face (visual input), and the produced information is aggregated by a higher-level module.

The design approach for each reasoning modules depends on what kind of environmental features is the subject of the reasoning. In some cases, where sensory information is not affected by noise, and the data fusion process can be easily coded, it is possible to choose a rule-based approach. On the contrary, if the reasoning module has to cope with uncertainty, as is the case where the goal is to detect user's presence, it is desirable that the design rely on the Bayesian Network theory, which allows to infer knowledge through a probabilistic process, and offers an effective way to deal with unpredictable

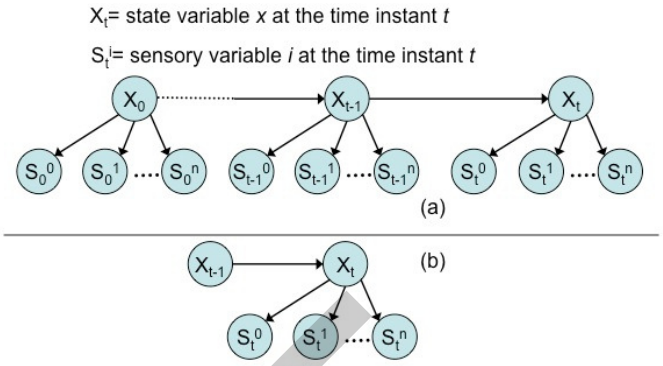

Fig. 3. Structure of a Markov chain for inference a given state feature starting from a set of sensory data.

ambiguities from multiple sensors [24]. This approach is different from a rule-based approach, that is not suitable for dealing with environmental features characterized by a large uncertainty, as the set of logical rules constituting logical reasoning engine is exclusively deterministic; our domain, on the other hand, requires the integration of intrinsically noisy sensory information that, moreover, can only provide partial observations of the system state.

Classical Bayesian networks [25], however, may only provide a static model for the environment, which would not be suitable for the proposed scenario; we therefore chose dynamic Bayesian networks or, more specifically, Markov chains to implement our models which thus allow for probabilistic reasoning on dynamic scenarios, where the estimate of the current system state depends not only on the istantaneous observations, but also on past states.

Figure 3 (a) shows our proposed Markov chain used to infer probabilistic knowledge on a given state feature starting from a set of sensory data. Each state feature affects a set of sensory readings (we indicate each sensor node with s^i), that can be considered the perceivable manifestation of that state. The link among the current state and its sensory manifestation is given by the probabilist sensor model $P(s_t^i|x_t)$. Moreover the current state depends on past state according to a state transition probability $P(x_t|x_{t-1})$.

The belief about the value of a state variable is the conditional probability with respect to all past states and the whole set of observation from the initial time to the current time. Due to the simplification introduced by the Markov assumption, the current state belief depends only on the past state and on current observations:

$$\begin{aligned} Bel(x_t) &= P(x_t|x_0, \ldots, x_{t-1}, s_0^0, s_0^1, \ldots, s_t^n) \\ &= P(x_t|x_{t-1}, s_t^0, \ldots s_t^n) \end{aligned} \quad (1)$$

According to the Bayesian Network structure, this joint probability can be factorized as follows:

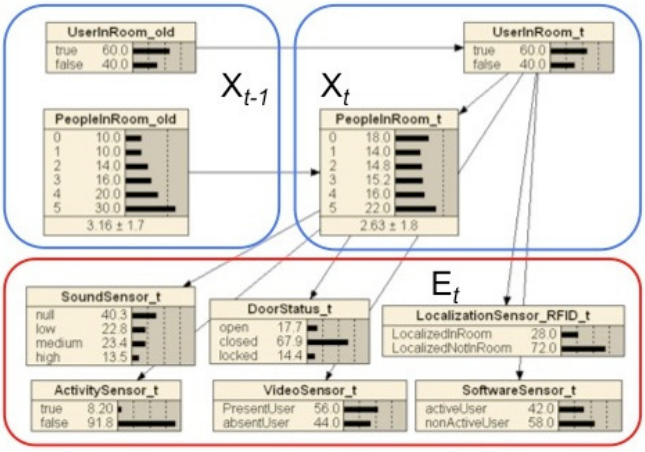

Fig. 4. Markov chain for room occupancy evaluation. X_t and X_{t-1} are the sets of state variables at the current and at the past time respectively, while E_t is the set of observable variables at the current time.

$$Bel(x_t) = \eta \left[\sum_{x_{t-1}} P(x_t|x_{t-1}) Bel(x_{t-1}) \right] \left[\prod_i P(s_t^i|x_t) \right] \quad (2)$$

Thanks to these simplifications, at each time step it is necessary to consider a reduced set of variables, as shown in Figure 3 (b), so reducing the number of required computation.

These principles had been followed in the design of the subsystem aimed at detecting the user's presence and therefore at reasoning on room occupancy. The outcome of this subsystem provides an estimate about the number of people present in the user's office room, and a probability for the user's presence as well.

Since there exist two interconnected state variables, i.e., two variables that are not probabilistically independent, the Bayesian network has been extended to manage two state variables, as shown in Figure 4. Sensor nodes are split into two sets, each of them is considered the measurable manifestation only of one hidden state variable. The two state variables are connected by dependency; in other words the number of people in the user's office room (associated to the *PeopleInRoom* variable) is influenced also by the presence of the considered user in their own office room (*UserInRoom*)

The state is observable through sensory information associated to the noise level in the room (*SoundSensor*), to the sensed interaction of the user with the room actuators (*ActivitySensor*), to the open / closed / locked status of the room door (*DoorStatus*), to the RFId-based naive user localization (*Localization-Sensor-RFId*), to the user's activity at their workstation monitored via software sensors (*Software-Sensor*), and to the video sensors (*VideoSensor*) Variables modeling this sensory information are connected with state variables through sensor probabilistic models, expressed by conditional probability tables that were learned from an opportune training dataset. Almost all of the above mentioned sensory information is discrete and does not require conceptual

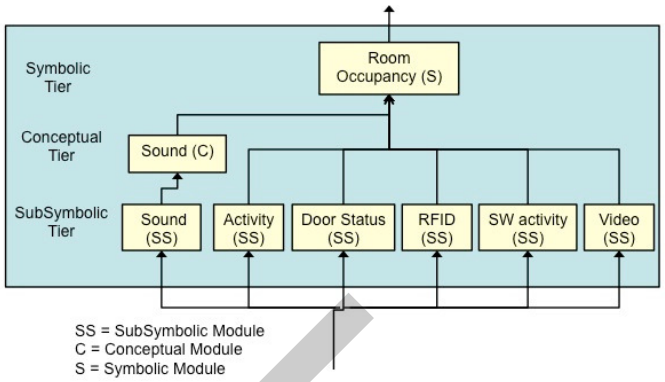

Fig. 5. The subsystems for lighting adequacy and room occupancy.

modules for extracting factual information from qualitative data, with the exception of the noise level, whose attached conceptual module uses a statistical characterization of room noise to classify it as `Negligible Noise`, `LowNoise`, `MediumNoise`, or `HighNoise`.

Figure 5 shows those architectural modules. The information outcome of the *Activity (SS)*, *DoorStatus (SS)*, *RFId (SS)*, *SW activity (SS)*, *Video (SS)* subsymbolic modules is directly handed over to the *Room Occupancy (S)* symbolic module that implements the previously described Bayesian network, while qualitative information produced by the subsymbolic module *Sound (SS)* needs preliminary classification through the *Sound (C)* conceptual module, before passing to the *Room Occupancy (S)* module.

IV. Case Study

The described architecture has been tested on a specific application scenario consisting in the management of an office environment, namely a university building, in order to fulfill constraints deriving both from the specific user's preferences about the air quality, and room lighting and occupancy, and from considerations on the overall energy consumption.

The sensory component of this system is implemented through a WMSN, whose nodes are equipped with off-the-shelf sensors for measuring such quantities as indoor and outdoor temperature, relative humidity, ambient light exposure and noise level. Sensor nodes have been deployed in various rooms close to "sensitive" indoor areas: by the door, by the window, and by the user's desk; additional nodes have also been installed on the building facade, close to the office windows, for monitoring outdoor temperature, relative humidity, and light exposure. Moreover, other nodes carry specific sensors, such as RFId readers, in order to perform basic access control. In our prototype, RFId tags have been embedded into ID badges for the department personnel, while RFId readers are installed close to the main entrance and to each office door; readings from each tag are collected via their coupled nodes, and forwarded by the WMSN to the intelligent core, that will process them and will reason about the presence of users in the different areas of the department. RFId-triggered reasoning about users' locations is inherently imprecise and

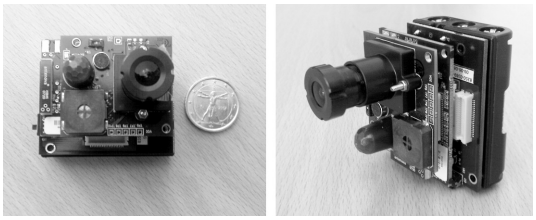

Fig. 6. Intel/Crossbow Imote2 with Imote2 Multimedia Board.

requires the integration with other sensory information, such as those collected by specialized software demons acting as virtual *software sensors* and used to detect the users' activity on their workstations. The users' interaction with actuators is also captured via ad-hoc sensor monitors. For instance, if the user manually triggers any of the provided actuators (e.g. the air conditioning, the motorized electric curtains, or the lighting systems) via the remote controls or traditional switches, specialized sensors capture the relative IR or electric signals so that the system may use them as implicit feedback.

The main contribution for detecting user's presence is given by video sensors integrated with wireless sensor nodes. Since Intel has developed several advanced wireless sensor node platform, we chose to develop our system using its state-of-the-art platform. The Imote2 is a smart device (36mm x 48mm x 9mm) produced by Crossbow and built around a low-power PXA271 XScale processor that can operate in a low voltage (0.85V), low frequency (13MHz) mode. It integrates an 802.15.4 radio with a built-in 2.4GHz antenna and can be expanded with extension boards.

In this work, we expanded it using the Imote2 Multimedia Board (IMB400) that integrates video and audio functionality into one platform (Fig. 6). In addition, the IMB400 features a Passive InfraRed (PIR) sensor to detect movement (up to 5 meters) for platform wake-up from sleep. Computer vision techniques have been developed using the Open Computer Vision (OpenCV) Library on the Imote2 Linux OS.

In order to validate our system, several tests were conducted on a prototype deployed for the monitoring of an office at the University of Palermo. Two Imote2 nodes were placed pointing at the entrance door of the office and to the user's desk respectively (Fig. 7), in order to monitor user presence.

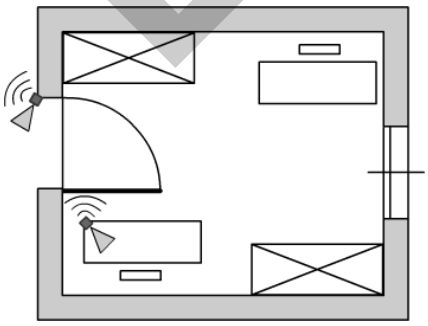

Fig. 7. Placement of the two Imote2 inside the office.

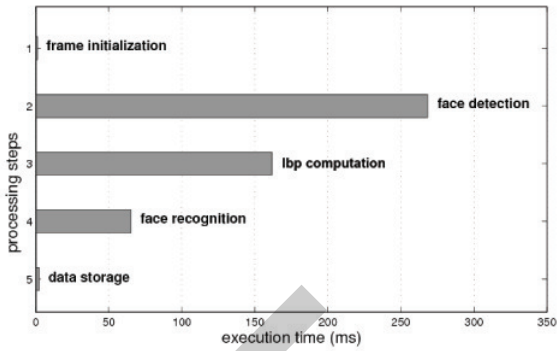

Fig. 8. Average execution time (ms) for each face processing step.

To evaluate the face recognition module we considered a scenario in which two nodes share the same face database. When processing a new frame, each detected face is described by using LBP, compared with already known faces and then the probability of match is computed.

We also tried to investigate the usage of each hardware component individually, but the producer does not provide benchmark tools to do it. Thus, additional tests have been performed to evaluate the overall system efficiency in terms of time of execution. Results are shown in Fig. 8.

Each bar represent the average execution time calculated for the corresponding face processing step. As you can see, face detection is the most computationally intensive operation since it requires several memory accesses for loading the training data and testing the whole input image. LBP computation and face recognition require $\sim$160ms-60ms respectively, while just a few milliseconds are required for frame initialization and final data storage.

To evaluate the whole data fusion process, we tested our system over a 3-days time period considering one user target unaware of the ongoing experiment, and so not modifying his usual behavior. To validate system results we compared them with videos obtained by the surveillance system already present at the Department.

The plots reported in Figure 9 graphically show how the system performs in detecting user's presence. To obtain a statistical evaluation of system performance we divided the considered time interval in a set of discrete time steps, in order to compute false positives and false negatives, and then the specificity and the sensitivity of the system, according to equations 3:

$$\begin{aligned} specificity &= \frac{\#\ of\ true\ negatives}{\#\ of\ true\ negatives + \#\ of\ false\ positives} \\ sensitivity &= \frac{\#\ of\ true\ positives}{\#\ of\ true\ positives + \#\ of\ false\ negatives} \end{aligned} \tag{3}$$

With respect to 300 considered time steps, the user's detection system produces 56 true positives, 234 true negatives, 6 false positives and 4 false negatives. Thus, we can conclude that the system shows an excellent behaviour with a specificity degree of 97,5% and a sensitivity degree of 93,3%.

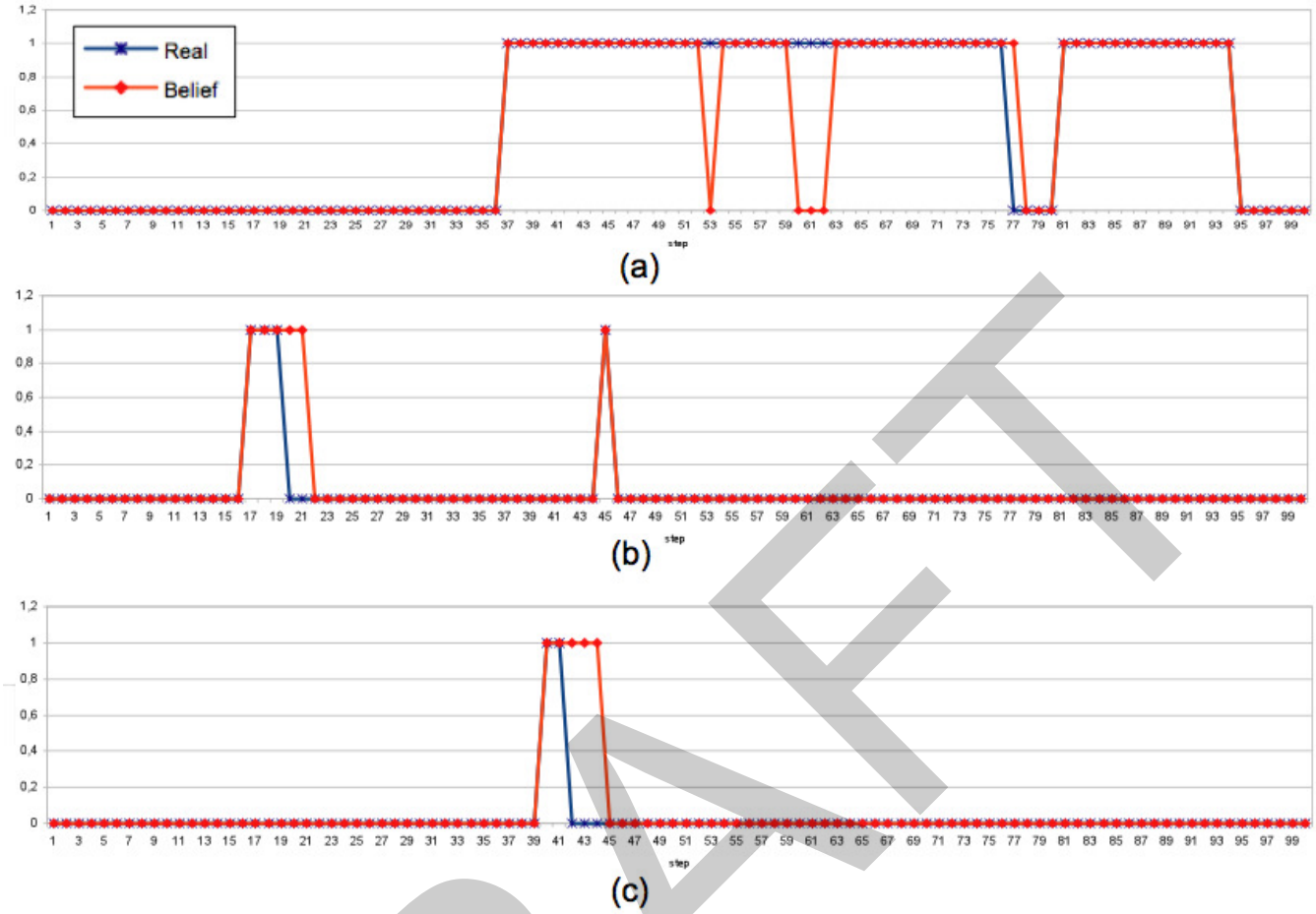

Fig. 9. Results of the bayesian multi-sensor data fusion process for user's presence detection.

V. Conclusion

This paper presented a multi-tier architecture for extracting higher-level knowledge from heterogeneous sensory measurements, also taking into account possible imprecisions in the original data, by means of a Bayesian approach. We focused on the specific issue of detecting the user's presence in selected locations of the monitored environment, in order to provide the grounds for subsequent reasoning in an Ambient Intelligence scenario. We also discussed the computer vision techniques employed for implementing face recognition on board of the nodes equipped with video sensors. Such information, together with additional information gathered by sensors of different nature (namely, RFId, audio sensors, and other activity sensors) was fed into our modular architecture in order to effectively implement sensor fusion, and to enrich the AmI system with context awareness. The provided experimental evaluation showed that the proposed Bayesian approach is promising in overcoming the difficulties arising from the inherently imprecision of sensory measurements, and allowed us to obtain a sufficiently precise estimate of the presence of the user.

References

[1] P. Remagnino and G. L. Foresti, "Ambient intelligence: A new multidisciplinary paradigm," *IEEE Trans. on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 35, no. 1, pp. 1–6, Jan 2005.

[2] K. Ducatel, M. Bogdanowicz, F. Scapolo, and J.-C. Burgelman, *Scenarios for Ambient Intelligence in 2010, Tech. Rep.* Seville: Information Soc. Technol., Advisory Group (ISTAG), Inst. Prospective Technol. Studies (IPTS), Feb 2001.

[3] D. Estrin, L. Girod, G. Pottie, and M. Srivastava, "Instrumenting the world with wireless sensor networks," in *Proc. of Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, Salt Lake City, Utah, May 2001.

[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–511–I–518 vol.1, 2001.

[5] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," *Proc. of European Conference on Computer Vision (ECCV 2004)*, pp. 469–481, 2004.

[6] I. Akyildiz, T. Melodia, and K. Chowdury, "Wireless multimedia sensor networks: A survey," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 14, no. 6, pp. 32–39, December 2007.

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

[8] 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*. Washington, DC, USA: IEEE Computer Society, 2003, p. 199.

[9] T. Yan, D. Ganesan, and R. Manmatha, "Distributed image search in camera sensor networks," in *SenSys '08: Proceedings of the 6th ACM conference on Embedded network sensor systems*. New York, NY, USA: ACM, 2008, pp. 155–168.

[10] D. Xie, T. Yan, D. Ganesan, and A. Hanson, "Design and implementation of a dual-camera wireless sensor network for object retrieval," in *IPSN '08: Proceedings of the 7th international conference on Information processing in sensor networks*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 469–480.

[11] K. Khattab, J. Mitteran, J. Dubois, and J. Matas, "Embedded system study for real time boosting based face detection," in *IEEE Industrial Electronics, IECON 2006 - 32nd Annual Conference on*, Nov. 2006, pp. 3461–3465.

[12] R. Kleihorst, M. Reuvers, B. Krose, and H. Broers, "A smart camera for face recognition," in *Image Processing, 2004. ICIP '04. 2004 International Conference on*, vol. 5, Oct. 2004, pp. 2849–2852 Vol. 5.

[13] L. Lo Presti, M. Morana, and M. La Cascia, "A data association approach to detect and organize people in personal photo collections," *Multimedia Tools and Applications*, pp. 1–32, Jul. 2011. [Online]. Available: http://dx.doi.org/10.1007/s11042-011-0839-5

[14] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," in *Proc. of International Conference on Analysis and Modeling of Faces and Gestures*. Springer-Verlag, 2007, pp. 168–182.

[15] B. Krishnamachari and S. Iyengar, "Distributed bayesian algorithms for fault-tolerant event region detection in wireless sensor networks," *IEEE Transactions on Computers*, vol. 53, no. 3, pp. 241–250, 2004.

[16] P. Castro and R. Muntz, "Managing context data for smart spaces," *IEEE Personal Communications*, vol. 7, no. 5, pp. 44–46, 2000.

[17] L. Marchesotti, S. Piva, and C. Regazzoni, "Structured context-analysis techniques in biologically inspired ambient-intelligence systems," *IEEE Trans. on Systems, Man and Cybernetics, Part A*, vol. 35, no. 1, pp. 106–120, Jan 2005.

[18] A. De Paola, A. Farruggia, S. Gaglio, G. Re, and M. Ortolani, "Exploiting the human factor in a wsn-based system for ambient intelligence," in *Complex, Intelligent and Software Intensive Systems, 2009. CISIS'09. International Conference on*. IEEE, 2009, pp. 748–753.

[19] A. Chella, M. Frixione, and S. Gaglio, "Understanding dynamic scenes," *Artificial Intelligence*, vol. 123, no. 1-2, pp. 89–132, 2000.

[20] E. Kandel, J. Schwartz, and T. Jessell, *Essential of Neural Science and Behavior*. New York: Appleton & Lange, 1995.

[21] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *EuroCOLT '95: Proceedings of the Second European Conference on Computational Learning Theory*. London, UK: Springer-Verlag, 1995, pp. 23–37.

[22] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 971–987, 2002.

[23] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," *Proc. of European Conference on Computer Vision (ECCV 2004)*, pp. 469–481, 2004.

[24] C. Lu, L. Fu, H. Meng, W. Yu, J. Lee, Y. Ha, M. Jang, J. Sohn, Y. Kwon, H. Ahn *et al.*, "Robust Location-Aware Activity Recognition Using Wireless Sensor Network in an Attentive Home," *IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING*, vol. 6, no. 4, pp. 598–609, 2009.

[25] R. Cowell, *Probabilistic networks and expert systems*. Springer, 1999.