UNIVERSITÀ
DEGLI STUDI
DI PALERMO

NDS
NETWORKING AND DISTRIBUTED SYSTEMS
RESEARCH GROUP

Mimicking biological mechanisms for sensory information fusion

Article

Accepted version

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

In Journal of Biologically Inspired Cognitive Architectures, Volume 3, January 2013, pp. 27-38.

It is advisable to refer to the publisher's version if you intend to cite from the work.

Publisher: Elsevier

http://www.sciencedirect.com/science/article/pii/S2212683X12000527

NDS LAB - Networking and Distributed Systems
http://www.dicgim.unipa.it/networks/

Mimicking Biological Mechanisms for Sensory Information Fusion

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

University of Palermo, viale delle Scienze, Ed. 6
{*alessandra.depaola, marco.lacascia, giuseppe.lore, marco.morana, marco.ortolani*}*@unipa.it*

Abstract

Current Artificial Intelligence systems are bound to become increasingly interconnected to their surrounding environment in the view of the newly rising Ambient Intelligence (AmI) perspective. This novel paradigm substantiates the need for context-aware reasoning, and calls for the deployment of pervasive and unobtrusive monitoring devices throughout the environment. In this paper, we present a comprehensive AmI framework for performing fusion of raw data perceived by sensors of different nature in order to extract higher-level information according to a model structured so as to resemble the perceptual signal processing occurring in the human nervous system. Following the guidelines of the greater BICA challenge, we selected the specific task of user presence detection in the system locality as a representative application clarifying the potentialities of cognitive models. Specifically, our contribution lies in the definition of a suitable model for knowledge representation and management; our goal is to make the artificial system able to *understand* the environment in which it acts, analogously to the way the human brain acts. In our system, the fusion of several information flows is performed by a Hidden Markov Model (HMM) that allows to deal with heterogeneous data, potentially affected by a non-negligible degree of uncertainty. Moreover, our approach allows to extract coherent concepts starting from a multimodal perception of the real world, also taking into account the history of past perceptions, in a computationally affordable way. Sensory data will be provided to the main inference engine after a preliminary processing performed by a wireless multimedia sensor network acting as a "peripheral nervous system".

Keywords: Information Fusion, Ambient Intelligence, Cognitive Architecture

Introduction

Today's advances in technology allow for the creation of cheap and unobtrusive monitoring devices that may be profitably used as a distributed sensory means permeating the whole environment. This allows for the immediate availability of huge amounts of raw sensory data which could easily flood and clog any high-level reasoning system, preventing it from proper functioning.

Many works have been presented in literature that address the issue of heterogeneous data analysis in order to provide a unitary representation at the symbolic level [1, 2], and the proposed solutions are typically very application-specific. A common trait, however, is the need for low-level pre-processing in order to refine data with increasing abstraction, so as to ease the

DRAFT

operation of an higher-level intelligent module devoted to ensure unitary reasoning on symbolic information.

In this work we focus in particular on the creation of an intelligent system for user detection through multi-sensor fusion in the context of an Ambient Intelligence scenario.

Ambient Intelligence (AmI) is a paradigm in Artificial Intelligence that introduces a shift in perspective as regards the role of the end user [3]. 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 [4]. In this respect, the basic intrinsic requirement of any AmI system is the presence of pervasive and unobtrusive sensory devices [5, 6], which are essential to ensure context-aware reasoning in order to act upon the environment, modify its state, and react to user-driven stimuli [7].

For the design of the interconnection of the high-level reasoning system with the low-level sensory infrastructure, we will follow an approach loosely inspired to the nervous system of complex biological organisms, that typically include some peripheral pre-processing mechanisms for extracting more significative information from the incoming wealth of sensory data. Striking examples may be found in some parts of the human nervous system, whose peripheral component deals with collecting sensory inputs, filtering them, and transferring them in an aggregated form to the central nervous system, where high-level processing will be performed.

The core component of our system, in particular, is inspired to the functional organization of human brain [8, 9], where different areas are functionally specialized for well-defined tasks for sensory signal processing. Besides functional specialization, also functional integration is performed in the different areas, and at different spatial and temporal scale. This suggested the design of a hierarchical and modular architecture, whose components operate independently and in parallel on different environmental stimuli in order to provide a symbolic representation of them. The interconnection among the different modules lets lower-level modules transfer their knowledge as input for higher-level ones that accept several simpler information streams and integrate them to provide a complex representation of the environment.

An early version of our architecture has been shown in [10]. In this paper we give a more detailed description of the AmI system and a few improvements will be presented for 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.

In this context, Wireless Sensor Networks (WSNs) represent an excellent choice for the underlying monitoring infrastructure, as they allow to get precise and continuous monitoring of the physical quantities of interest, as well as to perform basic *in-network* pre-processing of sensed data thanks to the limited computational capabilities of the nodes.

However, simple sensor nodes are not able to perceive high-level features such as *who* is in the office or *what* this person is doing there (e.g. reading, talking, using his/her 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 [11]) and face recognition (e.g., eigenfaces [12]) techniques have reached a good level of maturity, so we focused on their implemen-

DRAFT

tation 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 are sent to a server which will make inferences over the people interacting around each observed area. However, in a such dynamic scenario, sensory data are likely to be biased by environmental noise and by the unavoidably imperfect nature of sensor devices, so it is convenient to adopt an approach that is able to cope with uncertainty for developing reasoning components. Our architecture is thus designed to make use of a probabilistic approach that allows the overall system 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 detection of the user's presence and more in general the development of a full context awareness is a crucial functionality in many Ambient Intelligence applications. Some significative examples are the systems for assisting elderly people in their activities of daily living (ADLs) [13, 14], Building Management Systems (BMS) for minimizing the energy consumption in non-commercial buildings [15, 16, 17], and the systems devoted to the optimization of the environmental comfort in smart homes [18].

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

Related Work

Many works presented in Ambient Intelligence literature make use of WSNs both as a distributed sensory tool, and as a wireless network infrastructure. However, to our best knowledge, none of them fully exploits the potential computational capabilities of the sensor nodes; rather they are typically used as a mere data collection tool, with distributed sensors and communication capabilities.

In the work by Han, *et al.* [19], WSNs are used to provide inputs to an ambient robot system. Inside what the authors define a ubiquitous robotic space, a semantic representation is given to the information extracted from a WSN, but again this is used only as a sense-and-forward tool.

In [20], a WSN-based infrastructure is described targeting the development of wildfire prevention system, whose architecture is based on three layers, the lowest of which relies on a sensor network for measurement gathering.

Also the work presented in [21] employs a WSN, but the goal is the collection of information about the occupancy of the monitored premises; collected data are aggregated in order to compute predictions about the occupant behavior.

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* [22], are enabling several new applications such as multimedia surveillance sensor networks, environmental monitoring and many others.

Traditional wired distributed monitoring systems are deeply connected to their design constraints so that a reengineering process often requires a great effort making the system static and unmodifiable.

DRAFT

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.

Some works on embedded solutions for face detection or recognition have also been proposed. In [23] 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. In [24] the authors describe an architecture to perform real-time face recognition using smart cameras. The system consist of a cascade of filters for detection, registration and normalization and an RBF neural network for face recognition. A similar approach can be found in [25]. However many studies are usually limited to the implementation of a single face processing task.

Automatic face processing for recognition involves at least three different subtasks: *face detection*, *feature extraction*, *face recognition* and/or *verification*.

The face detection module we developed is based on Viola-Jones [11] face detector (VJFD), that is the most stable and used face detector both in academic and commercial systems. Face recognition step is based on a principal component analysis (PCA) technique: *eigenfaces* [12]. The reason for this choice is that *eigenfaces* is one of the most mature and investigated face recognition method and it performs well while normalizing faces with respect to scale, translation and rotation. Many others face recognition techniques have been proposed in literature. Two of the most common alternatives to eigenfaces are fisherfaces (based on Fisher Linear Discriminant Analysis) [26] and Local Binary Pattern (LBP) [27]. However, *eigenfaces* is more suitable for real-time analysis since it is mainly based on simple matrix operations. Moreover, while considering a limited set of different individuals eigenfaces results outperforms other methods [28].

To reduce the effects of different illumination conditions, we used the method presented in [29]. First a gamma correction is applied, then DoG (Difference of Gaussian) filtering is used for reducing shading effects. Finally, contrast equalization is performed to rescale the image intensities.

In the WMSNs literature a lot of works are presented about sensor fusion algorithms, in order to cope with natural uncertainty related to this kind of technology. In [30] a distributed bayesian algorithm is proposed to perform the sensor fusion task, and to solve the the fault-event disambiguation problem in sensor networks.

Ambient Intelligence applications generally exploit a wide range of sensory devices, so the sensor-fusion problem is an hot topic in this kind of research.

The idea of using video sensors as high-content sensors, in conjunction with other technologies, has been already expressed in the first phase of Smart Environment research [31], and in the following this idea confirmed its strength.

In such scenario several early works exploited only a conjunction of video and audio sensors [32, 33], while other more recent works exploit a more complex set of technologies.

Authors of [1] propose an AmI system able to process and analyze simultaneous data coming from a heterogeneous network of sensors (virtual sensors, CCD cameras, probes, etc.), combining them in a unique and symbolic representation of what happens in the monitored environment. This goal is achieved using a paradigm based on a neural fusion method that merges observations coming from state sensors, in order to extract context information.

In our work the fusion of information coming from different sensors, including RFId readers, ambient sensor and video sensors is made through a Bayesian Network in which each piece of

DRAFT

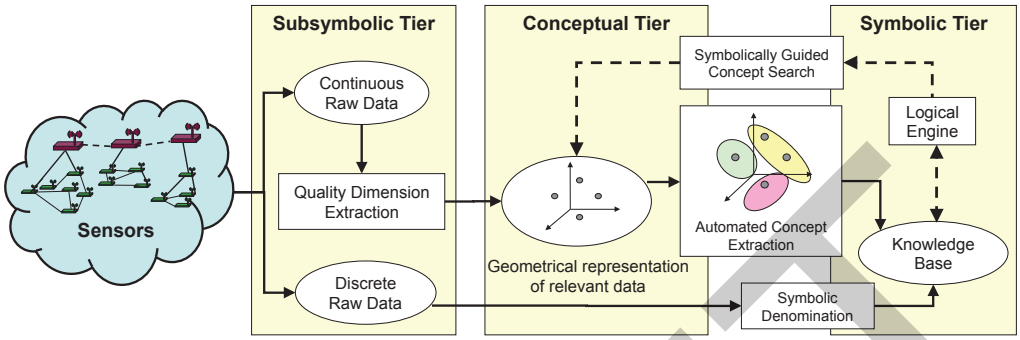

Figure 1: The three-tier structure of a low level module.

information is a signal generated by the corresponding subsystem. A similar approach can be found in [34], that presents a Kalman filtering method for unifying data coming from inertial sensors and computer vision with the goal of estimate position and orientation of camera in an augmented reality scenario. The image information obtained through a computer vision system is considered to be sensor signals in the Kalman filtering framework.

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 a 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.

Detecting the presence or absence of a user in a given environment is a fundamental functionality for the entire system; it thus cannot be delegated to simple low-level data processing, and rather requires more advanced reasoning capabilities. Traditional techniques of Artificial Intelligence usually show their limitations when employed with large amounts of sensorial data, so we propose here a multi-level cognitive architecture, where the process of knowledge extraction is carried on by several modules at increasing degrees of abstraction; this organization aims to gradually reduce the amount of data to be processed at each level, while increasing the information content of each information element.

The remote, distributed sensory device 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.

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.

DRAFT

Multi-tier Knowledge Representation

The proposed system is based on a multi-tier paradigm for performing knowledge extraction starting from sensory data. 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. This structure resembles the ideas presented in [35] that were applied to an artificial vision scenario; our system enhances this knowledge representation paradigm with the introduction of WMSNs as the lowest-level pervasive data acquisition means, and by reproducing the same 3-tier schema so that the abstract information extracted by the low-level modules of the architecture may be used as input for higher-level modules, thus producing more and more abstracted vision of the world surrounding the system itself.

The sub-symbolic 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 are 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 are outright handed over to the symbolic tier, where a linguistic representation will be given.

At the conceptual tier, data are 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 components 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 [35]: 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.

DRAFT

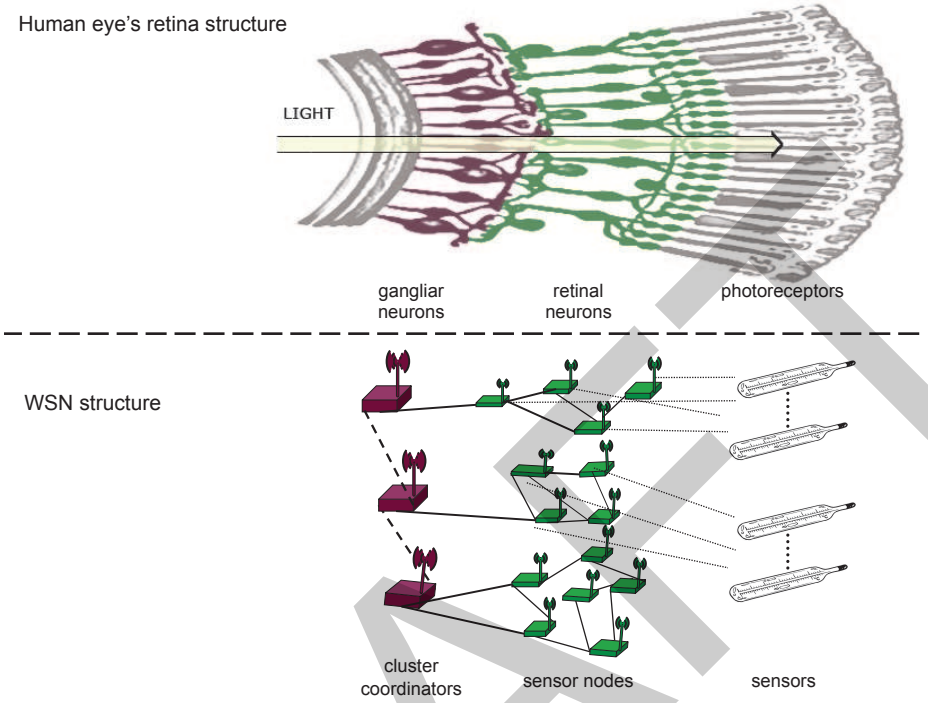

Figure 2: Comparison between the structures of the human retina and the proposed WMSN.

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 [9]: in the human eye, photoreceptors, although representing the innermost layer of the retina, convert light into electrical signals that are passed ahead to a network of retinal neurons, and are modified before being transmitted to ganglion 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. Figure 2 highlights the similarity between the structures of the human visual organ and of the WMSN employed here.

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 have 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

DRAFT

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.

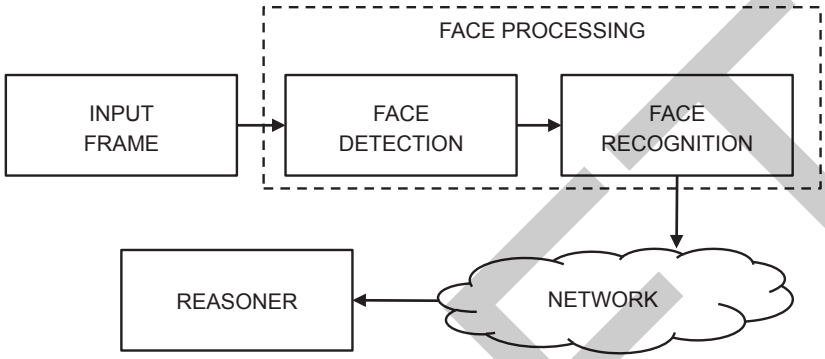

Figure 3: Scheme of the face processing module.

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 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. The *eigenfaces* is an information theory approach, based on Principal Component Analysis (PCA), to code and decode the information content of face images.

A two-dimensional $N \times N$ image can be considered as a point in a N^2-dimensional space, called image space. A set of different images then maps to a collection of points in this space, but face images will occupy a relatively low dimensional subspace. PCA provides a method to find the vectors that best represent the distribution of face images within the whole image space.

Let a training set X of M face images x_i of size $1 \times N^2$, the average face of X is defined by the vector $\bar{x}$:

$$\bar{x} = \frac{1}{M} \sum_{i=1}^{M} x_i, \tag{1}$$

and the difference between each face image x_i and $\bar{x}$ is stored in matrix A:

DRAFT

$$A = \left[\begin{array}{cccc} (x_1 - \bar{x}) & (x_2 - \bar{x}) & \cdots & (x_M - \bar{x}) \end{array} \right]$$
$$= \left[\begin{array}{cccc} \Phi_1 & \Phi_2 & \cdots & \Phi_M \end{array} \right]. \quad (2)$$

PCA then seeks the $M-1$ orthogonal vectors which best describe the distribution of the input data, which amounts to computing the eigenvectors of the covariance matrix C defined by:

$$C = \frac{1}{M} \sum_{i=1}^{M} \Phi_i \Phi_i^T$$
$$= AA^T. \quad (3)$$

Matrix C is $N^2 \times N^2$, thus eigenvectors and eigenvalues computation is impractical. However there are only $M-1$ meaningful (nonzero) eigenvectors v_k and they can be computed considering matrix $L = A^T A$ of size $M \times M$ instead of matrix C.

The linear combination of the M input face images forms the eigenface u_k:

$$u_k = \sum_{i=1}^{M} v_{ki} \Phi_i \qquad k = 1, \ldots, M. \quad (4)$$

In practice, while a lot of eigenfaces are required for accurate reconstruction of the image it has been observed that a smaller number of eigenfaces is sufficient for identification. Thus, the M' eigenvectors ($M' < M$) of matrix L are chosen as those with the largest associated eigenvalues.

A new face image x is projected into *face space* producing a vector of weights $\Omega^T = \left[\begin{array}{cccc} \omega_1 & \omega_2 & \cdots & \omega_M \end{array} \right]$ that describes the contribution of each eigenface in representing the input face image, where

$$\omega_k = u_k^T (x - \bar{x}). \quad (5)$$

Thus, the face descriptor is given by the vector Ω^T and a face is classified as belonging to the individual k with a probability p depending on the euclidian distance ε_k, where $\varepsilon_k^2 = \|(\Omega - \Omega_k)\|^2$ and Ω_k is the vector describing the kth individual. In particular, the probability p is obtained by sampling a gaussian distribution constructed by evaluating the distance values of a set of user's images from his representative vector Ω_k, and considering mean and variance of their statistical distribution.

Compared to [10], here we preferred PCA to LBP (Local Binary Pattern) since, as shown in Fig. 9, even if LBP recognition is faster than PCA, the whole recognition process (features detection and matching) is faster using Eigenfaces. Both techniques are described and compared in [36].

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

DRAFT

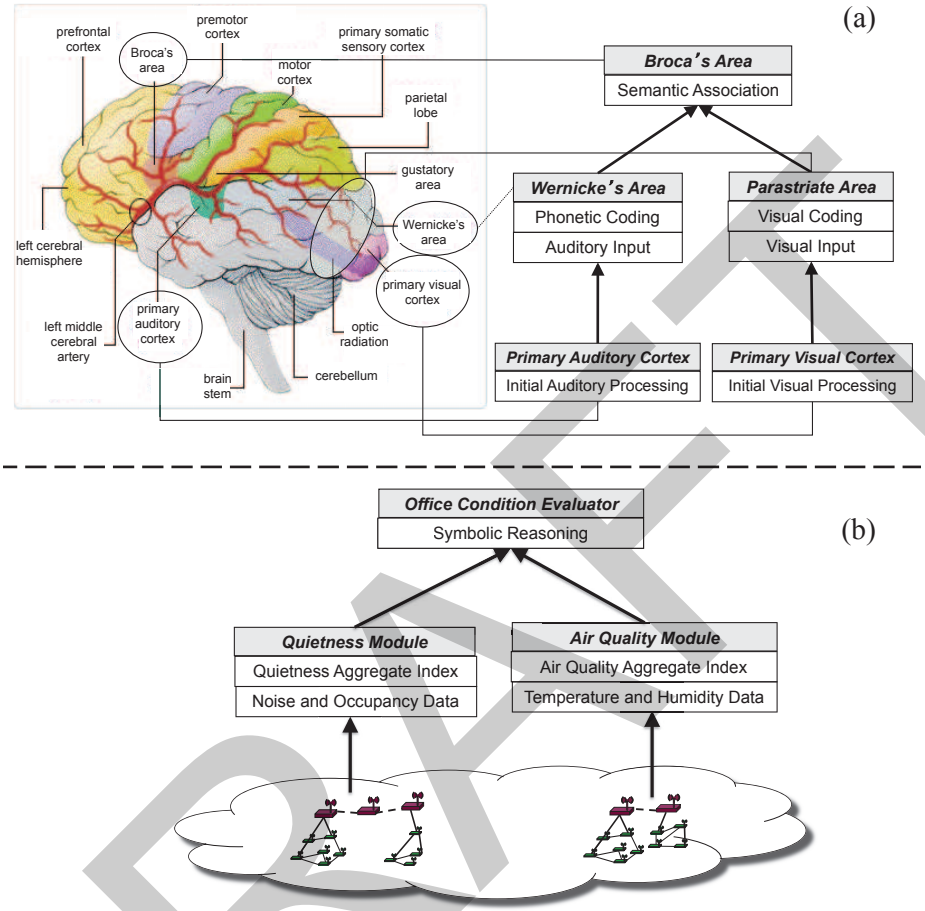

Figure 4: The human language comprehension model vs the proposed hierarchical reasoning model.

subsystems. From the design point of view, the modular organization allows for the realization of a scalable software architecture, able to effectively manage the huge amount of sensory data.

Figure 4 (partially taken from [1]) draws a parallel between the human brain model and our system model. 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 [9], provides a significant example of interaction patterns among specific areas of the brain, as schematically presented in the left side of Figure 4. 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.

[1] American Medical Association, http://braininfo.rprc.washington.edu

DRAFT

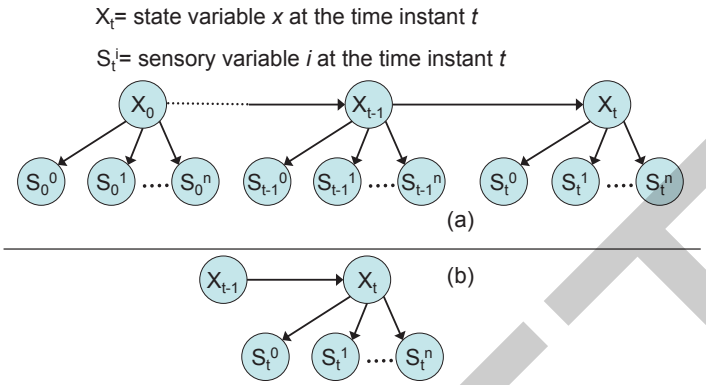

Figure 5: Structure of a Markov chain for inference a given state feature starting from a set of sensory data.

Pieces of information separately obtained by each low-level structure are sent to the areas devoted to phonetic and visual coding respectively. The outcome of the two intermediate modules are passed to the semantic association area, where they are merged.

In our architecture, an analogous example may be recognized in the modules devoted to detect user's presence, as shown in the right hand side of Figure 4. 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; thanks to a broader knowledge of the environment, it may perform more complex reasoning, without being overwhelmed by the incoming information thanks to the previous filtering.

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 ambiguities from multiple sensors [37]. This approach is unlike 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, 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 instantaneous observations, but also on past states.

Figure 5 (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

DRAFT

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:

$$\begin{aligned} Bel(x_t) &= P(x_t|x_0, \ldots, x_{t-1}, s_0^0, s_0^1, \ldots, s_t^n) \\ &= P(x_t|x_{t-1}, s_t^0, \ldots s_t^n). \end{aligned} \tag{6}$$

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

$$Bel(x_t) = \eta \left[\sum_{x_{t-1}} P(x_t|x_{t-1}) Bel(x_{t-1}) \right] \left[\prod_i P(s_t^i|x_t) \right]. \tag{7}$$

Thanks to these simplifications, at each time step it is necessary to consider a reduced set of variables, as shown in Figure 5 (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 are two interconnected state variables, that is two variables that are not probabilistically independent, the Bayesian network has been extended to manage two state variable, as shown in Figure 6. Sensory nodes are split into two sets, each of them is considered the measurable manifestation only of one hidden state variable. The two state variable are connected by dependency, that is the number of people in the user's office room (associated to the *PeopleInRoom* variable) is influenced also (there are some non measurable factor that influence this state variable) by the presence of the considered user in their own office room (*UserInRoom*). *UserInRoom* is a binary variable, while *PeopleInRoom* can take one among six values, corresponding to the number of people in the monitored room; namely the set of values is (`0`, `1`, `2`, `3`, `4`, `more then 4`).

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 (*SoftwareSensor*), 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 data set.

Almost all of the above mentioned sensory information is discrete and does not require conceptual 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`.

The evidence node corresponding to the *ActivitySensor* can take only two values (`true` or `false`, respectively) when there is some or none interaction between a user and an actuator, in the most recent time interval; while the evidence node that corresponds to the door status can take three values `open`, `closed` or `locked`. The *Localization-Sensor-RFId* and the *VideoSensor* evidence nodes are binary variables. The value of the former node depends on the output of a

DRAFT

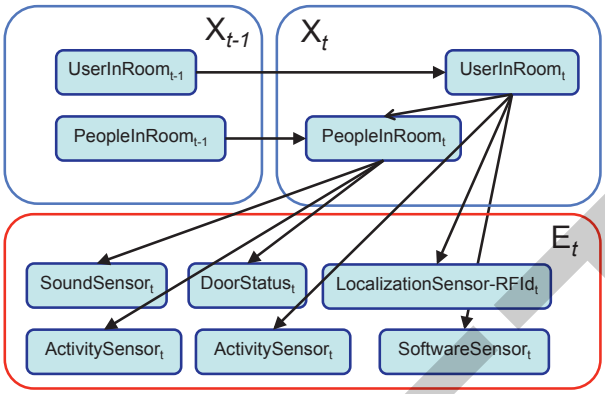

Figure 6: Markov chain for room occupancy evaluation. X_t and X_{t-1} are the sets of state variables at the current and at the past time respectively, while E_t is the set of observable variables at the current time.

virtual localization sensor described in the following Case Study section; while the value of the latter node depends on the face recognition process performed by the video sensors. Also the *SoftwareSensor* evidence node corresponds to a binary variable.

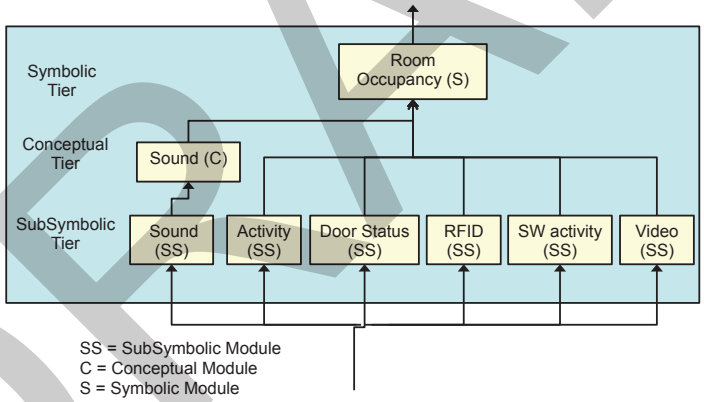

Figure 7: The subsystems for lighting adequacy and room occupancy.

Figure 7 shows those architectural modules. The information outcome of the *Activity (SS)*, *DoorStatus (SS)*, *RFId (SS)*, *SW activity (SS)*, *Video (SS)* sub-symbolic 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 sub-symbolic module *Sound (SS)* needs preliminary classification through the *Sound (C)* conceptual module, before passing to the *Room Occupancy (S)* module.

Results

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

DRAFT

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.

Case Study

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. The adopted communication architecture is based on the work presented in [38]. 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.

Some data produced by these simple nodes, namely temperature, humidity and light exposure measurements, are used as input of the AmI modules devoted to the automatic control of the environment conditions. Measurements of the sound level in the monitored environment can be exploited in order to obtain some information about the occupancy level of that room. If compared to video sensors, sound sensors provide a poor information, only partially correlated to the number of people in the room and not associated with the specific identity of the target user. Nevertheless gathering these data has almost no cost, and in a multi-data fusion process they can contribute with a piece of information, even if small.

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 readings are exploited by a virtual localization sensor, based on the use of Gaussian filters and the knowledge of the topology of the monitored building. In particular we used a simplified version of a Kalman filter [39], in which each sensory reading causes the change of the mean value of the Gaussian distribution that represents the belief about the location of the users. Since we don't know the direction of the movement of the user, if no further readings occur the uncertainty of the location belief increases. More details about this virtual localization sensor can be found in [38].

RFId-triggered reasoning about users' locations is inherently imprecise and 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.

DRAFT

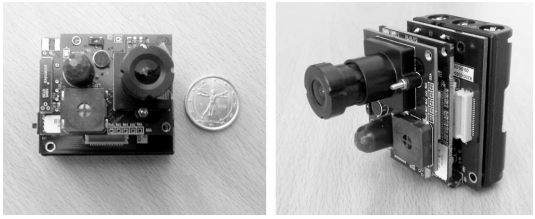

Figure 8: Intel/Crossbow Imote2 with Imote2 Multimedia Board.

In this work, we expanded it using the Imote2 Multimedia Board (IMB400) that integrates video and audio functionality into one platform (Fig. 8). 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 operating system [40].

Experimental Evaluation

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.

In particular, to evaluate the face recognition module we considered a scenario in which two Imote2 nodes share the same face space. Since our Department counts about 50 members, representing the set of known identities, we adopted a generic learning approach to estimate the face space offline by using a set of images taken from the Color FERET database[41]. When processing a new frame, each detected face is projected into the pre-computed eigenspace and the vector of weights that describes the contribution of each eigenface in representing the input frame is used as face descriptor.

The size of the descriptor we obtain for each input image is closely related to the value of M'. We performed several tests for evaluating PCA results while changing the number of training images and the size of each face. We finally considered a training set of $M = 200$ face images of size 61×61, previously aligned and cropped. The first $M - 1$ eigenfaces have been used for describing the faces in the personal collection.

We also tried to investigate the usage of each single hardware component, 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. 9.

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. The face representation and recognition processes using the eigenfaces approach require about ~100ms-120ms 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. Since we considered a public office, it usually happened that multiple people were in the observed room. Moreover the room is usually occupied by two persons and a number of users passing during the day, so that the use of a face recognition tool is mandatory. To validate

DRAFT

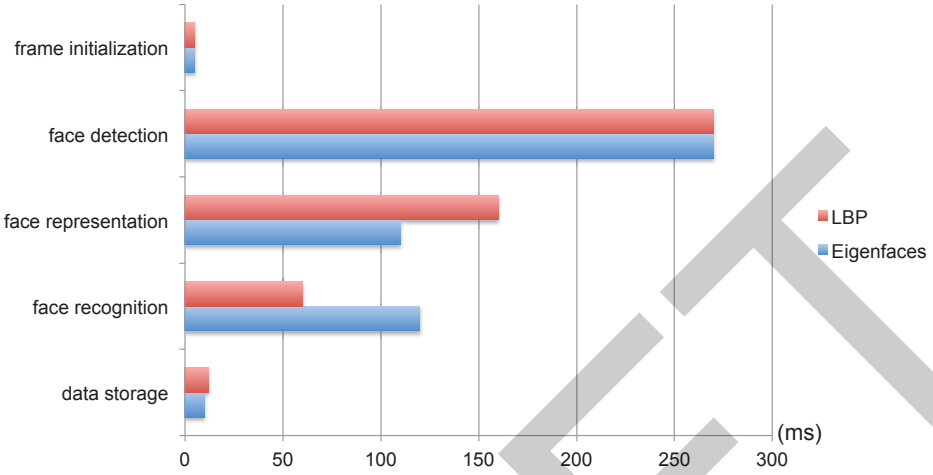

Figure 9: Average execution time (ms) for each face processing step using PCA (top) and LBP (bottom).

the system results we compared them with some videos captured by the surveillance system of the Department.

It is important to point out that our approach for performing a multi-modal sensory data fusion has been instantiated for the specific task of detecting the presence of a target user, and is to be regarded as a module of a more complex AmI system. For each user a module of this type runs in the central AmI server. The task of finding a trade-off between the requirements of different users has to be performed at a higher level, and it does not fall in the scope of this paper.

The conditional probability tables of the Bayesian networks are learnt on the basis of an evidence set collected over a 7-days time period and containing all the sensory events occurred in that period. The whole dataset has been manually annotated with the ground truth about both the number of people in the monitored room and the presence of the target user. Because the Bayesian network structure is a tree, learning the CPT was a straightforward process based on a frequentist approach. Thanks to the Markov Assumption, and again thanks to the tree structure, also the inference process is very simple and, above all, it is computationally affordable. The updating of the belief is performed according to the previous described Eq. 7.

To obtain a statistical evaluation of the performance of the multi-modal fusion performed by the Bayesian network 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 the following equation:

$$\begin{aligned} specificity &= \frac{\#\,of\ true\ negatives}{\#\,of\ true\ negatives + \#\,of\ false\ positives}, \\ sensitivity &= \frac{\#\,of\ true\ positives}{\#\,of\ true\ positives + \#\,of\ false\ negatives}. \end{aligned} \tag{8}$$

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 behavior with a specificity degree of 97,5% and a sensitivity degree of 93,3%.

DRAFT

Conclusion

In this paper we illustrated a biologically-inspired system for sensory information fusion. We focused on the specific issue of detecting users' presence in selected locations of an environment of everyday life, with the aim of providing the grounds for subsequent reasoning in an Ambient Intelligence scenario. The main contribution of the proposed approach is helpful for the development of context-aware artificial systems able to fully understanding the environment in which they act. Our proposal mimics the cognitive process that occurs in human mind when dealing with heterogeneous perceptual stimuli. In particular we discussed how visual information can be fused with other sensory inputs, through a hierarchical and modular BICA. The core of the system is constituted by a multi-tier architecture for extracting higher-level knowledge from sensory measurements, which takes into account possible imprecisions in the original data by means of a Hidden Markov Model (HMM). The HMM is fed with information flows produced by nodes of a wireless multimedia sensor network (WMSN) that preprocesses raw sensory data. In particular we discussed the computer vision techniques employed for implementing face recognition on board of special-purpose devices equipped with video sensors. The provided experimental evaluation showed that the adoption of a WMSN allows to obtain context-related sensory information with sufficient precision, if considered with respect to a single monitored feature. Moreover, we proved that, in an AmI scenario, the HMM approach allows to overcome the difficulties arising from the inherently imprecision of sensory measurements and produces a complex, but concise view of the current context whose global relevance is larger than the combined relevance of the constituting parts.

References

[1] L. Marchesotti, S. Piva, C. Regazzoni, Structured context-analysis techniques in biologically inspired ambient-intelligence systems, IEEE Trans. on Systems, Man and Cybernetics, Part A 35 (1) (2005) 106–120.

[2] C. Stauffer, W. Grimson, Learning patterns of activity using real-time tracking, IEEE Trans. on Pattern Analysis and Machine Intelligence 22 (8) (2000) 747–757.

[3] P. Remagnino, G. L. Foresti, Ambient intelligence: A new multidisciplinary paradigm, IEEE Trans. on Systems, Man, and Cybernetics–Part A: Systems and Humans 35 (1) (2005) 1–6.

[4] A. De Paola, A. Farruggia, S. Gaglio, G. Lo Re, M. Ortolani, Exploiting the human factor in a wsn-based system for ambient intelligence, in: Proc. of the International Conference on Complex, Intelligent and Software Intensive Systems, IEEE, 2009, pp. 748–753.

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

[6] M. Weiser, The computer for the 21st century, Scientific American 265 (3) (1991) 94–104.

[7] E. Aarts, H. Harwig, M. Schuurmans, Ambient Intelligence, The Invisible Future, J. Denning, ed., McGraw Hill, New York,, 2001.

[8] G. Tononi, G. Edelman, Consciousness and complexity, Science 282 (5395) (1998) 1846–1851.

[9] E. Kandel, J. Schwartz, T. Jessell, Essential of Neural Science and Behavior, Appleton & Lange, 1995.

[10] A. De Paola, M. La Cascia, G. Lo Re, M. Morana, M. Ortolani, User Detection through Multi-Sensor Fusion in an AmI Scenario, in: Proc. of the 15th International Conference on Information Fusion, Published by the IEEE Computer Society, 2012, pp. 2502–2509.

[11] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, pp. 511–518.

[12] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.

[13] A. Wood, G. Virone, T. Doan, Q. Cao, L. Selavo, Y. Wu, L. Fang, Z. He, S. Lin, J. Stankovic, Alarm-net: Wireless sensor networks for assisted-living and residential monitoring (2006).

[14] G. Virone, A. Sixsmith, Toward information systems for ambient assisted living, in: Proc. of the 6th International Conference of the International Society for Gerontechnology, 2008, pp. 1–4.

DRAFT

[15] J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, K. Whitehouse, The smart thermostat: using occupancy sensors to save energy in homes, in: Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems, SenSys '10, ACM, 2010, pp. 211–224.
[16] K. Kobayashi, M. Tsukahara, A. Tokumasu, K. Okuyama, K. Saitou, Y. Nakauchi, Ambient intelligence for energy conservation, in: Proceedings of the 2011 IEEE/SICE International Symposium on System Integration (SII), IEEE, 2011, pp. 375–380.
[17] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, T. Weng, Occupancy-driven energy management for smart building automation, in: Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building, ACM, 2010, pp. 1–6.
[18] J. Gil-Quijano, N. Sabouret, Prediction of humans' activity for learning the behaviors of electrical appliances in an intelligent ambient environment, in: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, Vol. 2, IEEE, 2010, pp. 283–286.
[19] K. Han, J. Lee, S. Na, W. You, An ambient robot system based on sensor network: Concept and contents of ubiquitous robotic space, in: Proc. of the Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2007, pp. 155–159.
[20] D. Stipanicev, L. Bodrozic, M. Stula, Environmental intelligence based on advanced sensor networks, in: Proc. of the Workshop on Systems, Signals and Image Processing, 2007, pp. 209–212.
[21] M. Akhlaghinia, A. Lotfi, C. Langensiepen, N. Sherkat, Occupant behaviour prediction in ambient intelligence computing environment, Journal of Uncertain Systems 2 (2) (2008) 85–100.
[22] I. Akyildiz, T. Melodia, K. Chowdury, Wireless multimedia sensor networks: A survey, Wireless Communications, IEEE 14 (6) (2007) 32–39.
[23] T. Yan, D. Ganesan, R. Manmatha, Distributed image search in camera sensor networks, in: Proc. of the 6th ACM conference on Embedded network sensor systems, 2008, pp. 155–168.
[24] R. Kleihorst, M. Reuvers, B. Krose, H. Broers, A smart camera for face recognition, in: Proc. of the International Conference on Image Processing, 2004, pp. 2849–2852.
[25] F. Zuo, P. de With, Real-time embedded face recognition for smart home, IEEE Trans. on Consumer Electronics 51 (1) (2005) 183–190.
[26] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.
[27] T. Ahonen, A. Hadid, M. Pietikäinen, Face recognition with local binary patterns, in: Proc. of European Conference on Computer Vision, Springer, 2004, pp. 469–481.
[28] A. Martinez, A. Kak, Pca versus lda, IEEE Trans. on Pattern Analysis and Machine Intelligence 23 (2) (2001) 228 –233.
[29] X. Tan, 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.
[30] B. Krishnamachari, S. Iyengar, Distributed bayesian algorithms for fault-tolerant event region detection in wireless sensor networks, IEEE Trans. on Computers 53 (3) (2004) 241–250.
[31] I. Essa, Ubiquitous sensing for smart and aware environments, Personal Communications, IEEE 7 (5) (2000) 47–49.
[32] I. Mikic, K. Huang, M. Trivedi, Activity monitoring and summarization for an intelligent meeting room, in: Proc. of the Workshop on Human Motion, 2000, IEEE, 2000, pp. 107–112.
[33] M. Trivedi, H. Kohsia, I. Mikic, Intelligent environments and active camera networks, in: Proc. of the IEEE International Conference on Systems, Man, and Cybernetics, 2000, Vol. 2, IEEE, 2000, pp. 804–809.
[34] J. Caarls, P. Jonker, S. Persa, Sensor fusion for augmented reality, in: Proc. of Ambient Intelligence: First European Symposium, Vol. 2875, Springer, 2003, pp. 160–176.
[35] A. Chella, M. Frixione, S. Gaglio, Understanding dynamic scenes, Journal of Artificial Intelligence 123 (1-2) (2000) 89–132.
[36] L. Lo Presti, M. Morana, M. La Cascia, A data association approach to detect and organize people in personal photo collections, Journal of Multimedia Tools and Applications (2011) 1–3210.1007/s11042-011-0839-5.
[37] 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 Trans. on Automation Science and Engineering 6 (4) (2009) 598–609.
[38] A. De Paola, S. Gaglio, G. Lo Re, M. Ortolani, Sensor9k: A testbed for designing and experimenting with WSN-based ambient intelligence applications, Pervasive and Mobile Computing. Elsevier 8 (3) (2012) 448–466.
[39] R. Kalman, A new approach to linear filtering and prediction problems, Journal of Basic Engineering 82 (1) (1960) 35–45.
[40] E. Ardizzone, M. La Cascia, M. Morana, Face processing on low-power devices, in: Proc. of the 4th International Conference on Embedded and Multimedia Computing, 2009, pp. 1–6.
[41] P. Phillips, H. Moon, S. Rizvi, P. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE

DRAFT

Trans. on Pattern Analysis and Machine Intelligence 22 (10) (2002) 1090–1104.

DRAFT