A MULTISENSOR SYSTEM FOR AMBIENT INTELLIGENCE APPLICATIONS

Andrea Turolla, Luca Marchesotti and Carlo Regazzoni

DIBE-University of Genova
Via dell'Opera Pia 11
16100 Genova,ITALY
carlo@dibe.unige.it

ABSTRACT

In this paper a complete architecture for Ambient Intelligence applications is presented. In particular the role of Data Fusion is shown to be performing to integrate heterogeneous information at different levels of abstraction. An Event Analysis method is presented in order to classify human activities within environments of interest whereas a location estimation method based on visual data is described in terms of logic functional architecture.

1. INTRODUCTION

Ambient intelligence is at the moment an open research field, still not bounded for what concerns topics of interest and related issues. Systems of this kind anyhow have as common aim the instantiation of "augmented" interactions with users by providing them with customized services. To achieve this, contextual information has to be gathered in order to get a high level description of the events taking place in an environment of interest. In literature, a certain number of integrated projects are in progress to show the potentiality of this technology in actual test beds. Among the others, a seminal work is presented by Trivedi, Huang e Mikic. It uses several cameras and microphones to acquire information on the surrounding environment [1]. More recently at the Artificial Intelligence Lab of the M.I.T. an *Intelligent Room* has been designed and developed. The research has been focused on natural and reliable vocal interfaces, context-specific user-transparent system reactions, dynamic resource allocation and natural cooperation among different contiguous environments [2]. Another important work aiming at developing strategies for context data extraction and management is represented by VICOM (Virtual Immersive COMmunication). The EU community provided in 2001 reference scenarios for Ambient Intelligence (AmI) [3] in order to highlight possible application fields. This raised a widespread interest on AmI with projects such as VICOM [4] or PER2 [5] which main goal is to integrate existing technologies in the field of Video Processing with advances in Cognitive Science domain. Other pioneers projects are Aura [6] in the field of distributed computing and oxygen [7]: the first proposes an innovative system for heterogeneous services which persists regardless of location, the latter develops an architecture enabling pervasive, human-centred computing through a combination of specific user and system technologies. Examples of this kind of systems go into the direction of applications able to integrate "awareness" (i.e. identification and tracking), "intelligence", and natural interaction. Our vision defines AmI systems as a set of virtual entities owning three fundamental capabilities: analysis, awareness and interaction. Considering the distributed nature of these kind of systems and the intrinsic heterogeneity of the handled information, a global approach exploiting Data Fusion techniques seems to be the most appropriate solution. This paper explicitly addresses these issues by proposing Fusion strategies to successfully combine heterogeneous data at different levels of abstraction. In particular, in section 2 an overview on the global structure of the system is reported, in section 3 the focus is given to the analysis capabilities and in section 4 to the fusion methodologies. Results and conclusions are reported in section 5 and 6.

2. THE LOGIC FUNCTIONAL ARCHITECTURE

The logic-functional structure of the proposed system is reported in **Figure 1**. As it can be seen, the complete architecture has been designed with a user-centred loop composed by four different clusters of modules. A set of heterogeneous sensors (i.e.: Sensing Cluster) collects data regarding objects of interest, which are typically represented by humans interacting in a given monitored environment such as an university laboratory where a set of different Computational Units (e.g.: Desktop PCs, PDAs) and communication facilities (i.e.: 802.11 WLAN) are tendered to students (i.e.: users of the AmI). A Video Analysis Module (VAM) processes multimedia data coming from a heterogeneous network of CDD-Video Cameras, whereas a Context Analysis Module (AMM) instantiates software agents devoted to collection of contextual information (e.g.: network load on computational units forming the AmIS, CPU load, temperature in rooms, etc.). A Data Fusion Module (DFM) completes the Analysis Cluster by fusing in a common representation video and contextual data. Next Cluster in the loop corresponds to the decisional step that has to translate on future actions/communications towards the user the information already collected and analysed. In addition, this information is matched against past decisions which have been previously stored in appropriate Databases (i.e.: Long Term Memory Db, User Db). The last cluster (i.e.: Action and Communication Cluster) represents the front-end of the system to users; in particular it has to render the information with multimodal and scalable communication methodologies (i.e.: animated avatars [8], sms, ect.) and to manage system actuator devices (i.e.: thermal regulators, door opening, light switches, etc.)

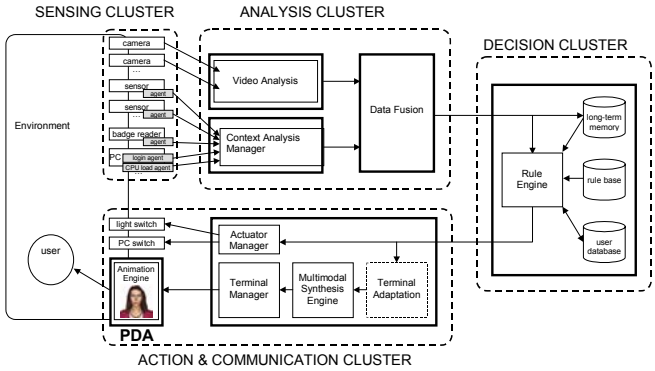


Figure 1. The overall logic-functional architecture for the system.

3. THE ANALISYS CLUSTER

The analysis cluster is mainly composed by three interconnected modules: the Video Analysis Module (VAM), the Context Analysis Module (CAM) and the Data Fusion Module that finally integrates the processed information in a parametric representation of events. In this scenario, VAM has been specifically designed to classify objects (e.g.: pedestrians, vehicles, others) interacting in the monitored environment and to extract information regarding their physical location. The module has been structured with a chain of logical tasks [9]: the first step is a Dynamic Change Detection performing the difference between the current image and a reference one (i.e. background). Each moving area (called Blob) detected in the scene is bounded by a rectangle to which a numerical label is assigned. Thanks to the detection of temporal correspondences among bounding boxes, a graph-based temporal representation of the dynamics of the image primitives can be built. The core part of such systems is however represented by the tracker algorithm that projects in a 2-D map location of objects through Camera Calibration Parameters [10]. The information regarding m-th object extracted at frame *k* from *i-th* Video Sensor is inserted in a feature vector referred to as Detection Report (DR):

$$\bar{d}^{i}_{k,m} = [\bar{p}^{i}_{m}(k), \bar{v}^{i}_{m}(k), \bar{c}^{i}_{m}(k)]$$

where $\bar{p}, \bar{v}, \bar{c}$ components respectively represents the position, speed and class of m-th object. Along with VAM a Context Manager Module (CMM) coordinates a society of Software Agents which migrates within system's computational units (i.e.: CU) and collects information regarding:

- all logins into the given PC, x^{u}_{AL}
- the network load, x^{u}_{NL}
- the CPU load $x^{u}_{CPU,L}$

All these parameters are estimated for each CU and inserted in a Contextual Report:

$$\bar{d}^{u}_{k} = \left[x^{u}_{AL}, x^{u}_{NL}, x^{u}_{CL}\right]$$

4. DATA FUSION MODULE

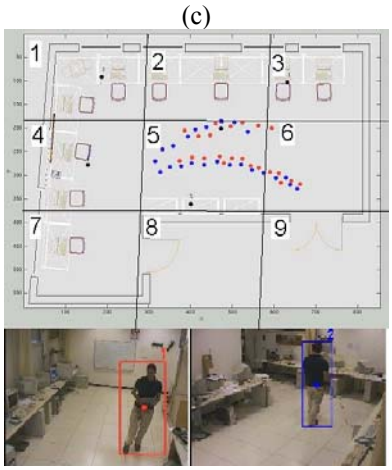
4.1 Introduction

The Data Fusion Module structure is reported in Fig. 3; as it can be seen, fusion is performed at two levels of abstraction (i.e.: Low Level Data Fusion and High Level Data Fusion) in order to fuse Detection Reports coming from different video cameras and Contextual Reports produced by Agents. In the Video Integrator three main steps are performed following the general fusion model proposed in [11]:

- *Data Alignment (DA)*
- *Data Association (DS)*
- *State Estimation (SE)*

DA initially synchronize in time and space DRs (Detection Reports) coming from the network of cameras whereas Spatial alignment is achieved through Camera Calibration based on classic Tsai method [10]. In the presented system, all video sensors have been calibrated with a common calibration strategy on a reference 2-D map.

Next step is represented by Data Association: once objects can be synchronously placed in a common 2-D space, a situation as the one depicted in **Figure 2** a-b has to be faced; two different views of the same scene are reported showing a moving object (i.e.: user of the lab). As it can be seen positions $\bar{p}^{i}_{m}(k)$ for the object are projected in the map and have to be associated into a unique and consistent track (i.e.: trajectory of object).



(a) **(b)**

Figure 2. (a)-(b)FOVs for the two cameras. (c) Laboratory map with projected tracks.

Relevant features for DA can be found at different levels:

- **Signal (Pixel) Level**: colour histograms
- **Object (Blob) Level**: shape, position

Different association metrics are used exploiting features of Detection Reports (i.e.: *speed*, *position* and *color*). The final output takes the form of a subset of DRs, which are

associated to a single track. Track's State Estimation represents last step towards fusion; in particular features of the track have to be updated taking into account the newly associated DRs. The principal feature to be estimated is anyhow the position of each track. Therefore, given the set of available DRs associated with the track, position is estimated using different approaches (e.g.: Generalized Hough Transform based method [12]) in relation to the status of the detected object that can be clear or occluded by other static or dynamic objects. Once tracks' state has been successfully estimated the following quantity is calculated:

$$x_P = \sum_{g=1}^{G} k_g n_g$$

x_P represents the distribution of objects with respect to the monitored environment. It is a feature calculated by subdividing the monitored region (e.g.: map in fig.2-c) in g parts (i.e.: with $1 < g \leq G$ G = total number of regions). The number of objects present in each region is evaluated by exploring the list of detected tracks. The resulting feature vector x_P is composed by the number n_g of objects (i.e.: tracks) present in each zone *g*, weighted by the *time of presence* k_g in g-th zone. Global motion is then evaluated from the mean speed of the objects moving in the scene:

$$x_v = \frac{1}{N_{TOT}} \sum_{i=1}^{N_{TOT}} \sqrt{v_{i_x}^2 + v_{i_y}^2}$$

where v_{i_x} and v_{i_y} are the *x* and *y* components of the speed of the *i*-th object in the scene.
A global vector is then assembled:

$$x_R = \left[x_v, x_p\right]$$

A similar vector is built within the Context Integrator starting from Context Reports $\bar{d}_k^u$; in this case the fusion process is fairly simple because the features for Network and CPU Load representative of the actual state of the architecture are simply averaged over the total number of Computational Units U:

$$x_C = [\, x_{AL}, x_{NL}, x_{CL} \,]$$

where:

$$x_{AL} = \sum_{0}^{U} x_{AL}^u,\ x_{NL} = \sum_{0}^{U} x_{NL}^u,\ x_{CL} = \sum_{0}^{U} x_{CL}^u$$

To this end a global feature vector integrating the MVI and the CI contributions can be assembled:

$$\bar{x}_E = \left[\bar{x}_R, \bar{x}_C\right] = \left[x_v, x_p, x_{AL}, x_{NL}, x_{CL}\right]$$

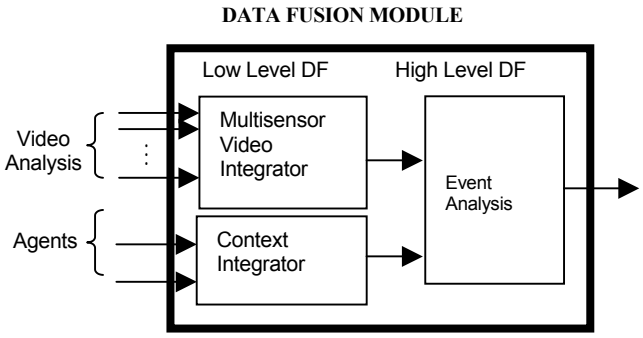


Figure 3. Data Fusion Module

4.2 Event Analysis

Event Analysis Fusion scheme (High Level Data Fusion) has the primary objective to determine the current event taking place in the region of interest by inspecting the integrated vector $\bar{x}_E$. The events to be discriminated are mainly represented by activities of people detected in the monitored area. In particular, seven different events have been taken into account: four of them specifically addressed to the description of a prolonged action (i.e.: WHL, high human work, WHF few human work, WAL few automatic work and WAF high automatic work) the other three a more instantaneous happening (i.e.: ARRIVE, arrive of new person, NULL, system not active and WAIT, nothing relevant). Self-Organizing Maps (SOM) [13] have been exploited as event classification method; they perform a spatial organization process of the input features (i.e.: vector $\bar{x}_E$) called *Feature Mapping* through an unsupervised learning technique (input and output are not mapped by an external supervisor). All required operations can be divided in six steps:

1. Map Initialisation: reference vectors are initialised to random values bounded from the minimum and the maximum of the learning set.
2. Map Training: a first phase called *ordering phase* in which the reference vector of the map unit are ordered and a second one where a fine-tuning is performed.
3. Evaluation of Quantization Error using different features vectors.
4. Map Calibration: the map units are calibrated using know input data sample defined by a label.
5. Map Visualization.
6. Real time work.

Training session has been performed by collecting 3769 features vectors in an offline mode. The data collection has been performed recording information from sensors for a total period of ten hours; this period has been divided into three subsets to get a sequence in early morning, half day and evening. This procedure has been performed in order to present to the system all the meaningful situations. The whole sequence has been reiterated to have 5000 input in the ordering phase and 200000 for the fine-tuned one. In the map calibration in which 680 features vectors, derived from some

particular situations, have been collected. The result of the training and calibration is the map layout shown in fig. 4.

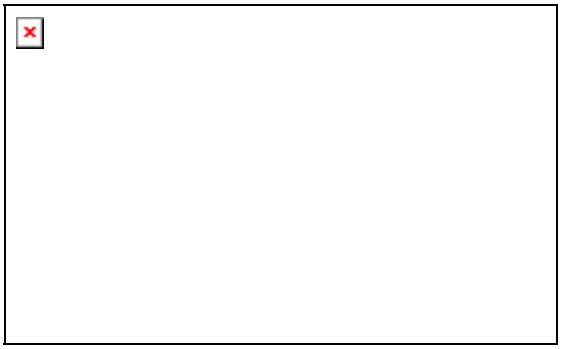


Figure 4. Self Organizing Map layout different neuron clusters are highlighted.

5. RESULTS

Results here proposed mainly regards High and Low Level Fusion Schemes. In particular correct association to tracks is evaluated in terms of confusion matrixes. Two situations have been tested, the first one (Table 1-(a)) reproduces the case in which all features vectors in DR are taken into account and 4 different tracks are present in the scene. As it can be seen, the displacement over the diagonal is null, meaning that there's no ambiguity on association whereas if speed feature is neglected in association, spreading outside the diagonal is evident and association is corrupted.

	t 1	t 2	t 3	t 4
t 1	218	0	0	0
t 2	0	218	0	0
t 3	0	0	187	0
t 4	0	0	0	218

(a)

	t 1	t 2	t 3	t 4
t 1	207	5	0	0
t 2	90	121	0	0
t 3	0	0	118	98
t 4	0	0	20	195

(b)

Table 1 Four trace with complete features vector (a) and () without Id in the feature vector.

For Event Analysis, proposed results derives from on-line tests carried out in real conditions. The figure 4 represents the U-matrix after the learning phase where it's possible to identify several clusters related to events. A ground-truth data set has been designed in order to quantitatively test the goodness of the trained map. In particular plots have been written annotating a sequence of actions, which an actor had to perform, and the time of the event. Results have been summarized in table 2, where a percentage of false alarm P_{FA} and correct decision P_D is reported for the recognition of the correct event.

6. CONCLUSIONS

A complete architecture for Ambient Intelligence applications has been presented. In particular a data fusion approach has been shown to be performing both for integrating low level and high level data. Results on correct classification for events of interest are satisfying; anyhow they can be improved with a more exhaustive training session and with refined features vectors.

7. ACKNOLEDGMENTS

This work was partially supported by the University and Scientific Research Ministry (MIUR) of the Italian Government under the National Interest Scientific Research Program and by ELSAG spa.

REFERENCES

[1] M. Trivedi, K. Huang and I. Mikic. Intelligent Environments and Active Camera Networks, University of California, San Diego, IEEE Transactions on Systems Man and Cybernetics, October 2000.

[2] Brooks, R. A. with contributions from M. Coen, D. Dang, J. DeBonet, J. Kramer, T. Lozano-Perez, J. Mellor, P. Pook, C. Stauffer, L. Stein, M. Torrance and M. Wessler, "The Intelligent Room Project" , Proceedings of the Second International Cognitive Technology Conference (CT'97), Aizu, Japan, August 1997.

[3] ISTAG; Scenarios for Ambient Intelligence in 2010; Final report, Feb 2001, EC 2001:http://www.cordis.lu/ist/istag.htm.

[4] http://vicom-project.it

[5] http://spt.dibe.unige.it/ISIP/Projects/miur02/main.html

[6] http://www-2.cs.cmu.edu/~aura/

[7] http://oxygen.lcs.mit.edu/

[8] C.Bonamico, C.Braccini, M.Costa, F.Lavagetto and R.Pockaj, "Using MPEG-4 parameters for calibration/animation of 3D Talking Heads", *Proc. of IWDC'02 - Tyrrenian International Workshop on Digital Communications* , Capri, Italy, September 8th - 11th, 2002, pp. 199-205

[9] L. Marcenaro, F. Oberti, G.L. Foresti and C.S. Regazzoni, "Distributed architectures and logical task decomposition in multimedia surveillance systems", Proceedings of the IEEE, Vol.89, N.10, Oct.. 2001, pp. 1355 –1367.

[10] Tsai, Roger Y., 1987, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses.", *IEEE Journal of Robotics and Automation* RA-3(4): 323-344, August 1987

[11] E. Waltz and J. Llinas, *"Multisensor data fusion"* , ISBN 0-89006-277-3, 1990 Artech House, Inc.

[12] *F.Oberti and C.Regazzoni*, **"Real-Time Robust Detection of Moving Objects in Cluttered Scenes"**, European Signal Processing Conference, Eusipco 2000, Tampere, Finland.

[13] T. Kohonen, J.Hynninen, J.Kangas, J.Laaksonen, Self-Organizing Map Program Package, University of Technology – Laboratory of Computer and Information Science, Helsinki, Finland, April 1995 .

Super-State	Real World	P_D	P_{FA}
WHF	Low Human Work	60%	40%
WHL	High Human Work	60%	40%
WAF	Low Machine Work	80%	30%
WAL	High Machine Work	80%	30%
ARRIVE	Laboratory Incomes	72%	50%
WAIT	Sleeping	70%	35%
NULL	Everything Stopped	98%	1%

Table 2. Results on Events Classification.