Extraction of Aligned Video and Radio Information for Identity and Location Estimation in Surveillance Systems

Luca Marchesotti, Reetu Singh and Carlo Regazzoni
DIBE-University of Genova
Via dell'Opera Pia 11
16100 Genova,ITALY
marchesotti@dibe.unige.it

Abstract– In this paper, an innovative schema towards fusion of Video and Radio Sensors is reported; in particular the focus is given to preliminary steps for integration namely information extraction and alignment. The final aim is to propose methodologies within the context of Surveillance Systems to successfully locate objects of potential interest (i.e.: humans) and identify them using compatible features extracted from Radio and Video Sensors.

Keywords: Data Alignment, Video Tracking, Radio Devices, Data Fusion, Surveillance Systems.

1 Introduction

Now-a-days there is a growing need in our society to guarantee a satisfying level of security in social environments. For this reason, the problem of surveillance has received growing attention in the last few years. The actual trend is to go in the direction of automatic or semi-automatic multisensor Surveillance Systems (SS). A SS can be defined as a technological tool that assists operators by providing an extended perception and reasoning capability about situations occurring in the monitored environment. However, there is a further scope to enhance the efficiency and capability of the SS other than the existing one. The Video Sensors as a part of the SS perform fairly well in objects location estimation whereas the video identification is still a challenge because of its operating in unconstrained environments (e.g.: parking lots, airports, etc.). Within this context, the successful location of objects of potential interest (i.e.: humans) and their identification are still open issues in the state of the art; anyhow different positioning approaches are available based on: Video [9] and Radio Signals [2].
The in abundance availibility of the radio devices in the market has given an opportunity to improve the functionality of SS as most of the moving objects are equipped with such devices. If the objects communicate with the transmitter through mobile devices where transmitter themselves are a part of the SS, in that case the identification of users are not a problem for Radio-Based Systems (RBS) on contrary to the Video-Based ones (VBS). The price to pay is however a less accurate localization data for RBSs. A tradeoff solution is clearly a system able to integrate heterogeneous sensors to perform localization and identification simultaneously. Following this aim, an architecture exploiting CCD-Video Cameras and 802.11 Wireless LAN positioning methodology is presented; in particular the paper specifically addresses and exploit one of the first step towards fusion that is called as Data Aligment, according to a popular Data Fusion model [1]. In this paper the main attention has been paid on the extraction of aligned video and radio information for further fusion investigation.
The paper layout follows: section II describes the system in terms of logic functional architecture whereas section III explores the alignment and extraction methods for both the radio and video case. Preliminary qualitative results collected during a system trial session are shown in section IV and conclusions are drawn in section V.

2 The Logic Functional Architecture

2.1 The formalism

The formalism hereinafter used to describe the logic functional architecture of the proposed systems assumes that a set of heterogeneous sensors $S = \{\overline{S^c} : c = 1,..,N_S\}$ is divided in N_S different classes $\overline{S^c} = \{s^c{}_i : i = 1,..,N_{\overline{S^c}}\}$ where $N_{\overline{S^c}}$ is equal to the number of sensors in class (c^{th}). Each sensor is directly connected to a dedicated Computational Units (i.e.: CU) belonging to the set $U = \{u_l : l = 1,..,N\mu\}$ with $N\mu$ equal to the total number of corresponding sensors. Each CU acquires data providing *Object Reports* (OR) $\bar{r}^c_{i,m}(k)$ for each object *m* found at time *k* from i^{th}.sensors in c^{th} class. OR is represented as a multidimensional vector composed by different features related to the detected object:

$$\bar{r}^c_{i,m}(k) = [\overline{f}^i_1(k),\ldots,\overline{f}^i_{N_r}(k)] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

with N_r the total number of features $\overline{f}$ in the report.
For each detected physical object tracks are instantiated and updated:

$$T_m(k) = \{\hat{r}_m(K-k) : k = 0,...,K\} \ldots\ldots\ldots\ldots\ldots (2)$$

with *K* = current time, m = detected object

Tracks are sequences of *estimated* reports $\hat{r}_m(i)$ derived from integration of heterogeneous ORs:

$$\hat{r}_m(k) = [\hat{f}_1^i(k), \ldots, \hat{f}_{N_r}^i(k)] \quad \text{.....(3)}$$

2.2 Architecture Layout

In Figure 1 the overall logic functional architecture of the proposed system is depicted. As as it can be seen the structure is inspired by a classical model of Data Fusion systems described in [1]. In this model three different levels of analysis have to be performed towards instantiation of fused tracks:

1. Data Aligment
2. Data Association
3. State Estimation

In the presented case, only two classes of Sensors have been included in the architecture: Static CCD Video Cameras and 802.11 WLAN Base Stations (i.e.: classes $c = R, V$). Data collected by sensors have to be aligned in order to be succesfully compared; two modules have been inserted in order to independently pre-process the informations

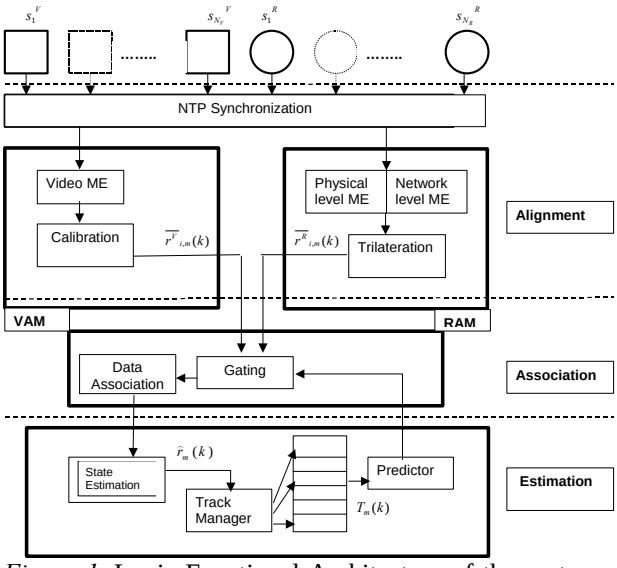

Figure 1. Logic Functional Architecture of the system composed by Data Aligment, Data Association and State Estimation

coming from the different sensors in dedicated CUs (i.e.: standard Pentium-based desktop PCs). In particular, a Video Analisys Module (i.e.: VAM) takes care of extracting metadata from video sources whereas the Radio Analisys Module (i.e.: RAM) does the same with Base Stations. Output Object Reports are respectively addressed as Video Object Reports (VOR) and Radio Object Reports (ROR) $\bar{r}_{i,m}^{c=V,R}(k)$. They have to be associatiated in relation to their features $\hat{f}^i(k)$ and according to specific association rules. Once different groups of reports are evaluated by the Data Association submodule, they have to be fused in estimated reports $\hat{r}_m(k)$ by the State Estimator (Figure 1). Ultimately, tracks $T_m(k)$ are updated or newly instantiatiated by the Track Manager whereas a Prediction submodule feedbacks future values for ORs into the Association step in order to get a closed –loop analisys. A more detailed treatement of the general architecture can be found in [10].

3 Alignment and Extraction of Object Reports

3.1 Introduction

In order to extract, associate and fuse data coming from multiple heterogeneous sensors the first operation to be carried out is Data Alignment (DA); this step can be further subdivided in Temporal Alignment (TA) and Spatial Alignment (SA). Whereas SA step is related to the typology of signals, TA is more concerned with the physical acquisition of signals and hardware set-up of the architecture. For this reason TA has been implemented in a common layer for all sensors belonging to S (i.e.: set of all available sensors) whereas an independent spatial alignment strategy has been designed for the two classes of sensors $c = R, V$ in order to project location data on a common reference plane. As it can be seen from Figure 1, dedicated submodules (i.e.: Video ME, Physical Level ME and Network Level ME) are specifically devoted to the extraction of metadata that is coded under the form of an Object Report:

$$\bar{r}_{i,m}^{c}(k) = [\bar{p}^i(k), \bar{id}^i(k), \bar{c}^i(k)] \quad \text{.....(4)}$$

where $\bar{p}^i(k), \bar{id}^i(k), \bar{c}^i(k)$ respectively indicates position, id and class (e.g.: pedestrian, vehicle, others) of detected objects moving in the monitored environment.

3.2 Temporal Aligment

In the presented case, acquisition devices for data collection are Video Frame Grabbers for A/D conversion and Base Station Traffic Analysers for recording WLAN signal power; these devices run on dedicated CUs (i.e.: Computational Units, standard Pentium based desktop PCs) and they have to be synchronized to get time aligned ORs $\bar{r}_{i,m}^{c=V,R}(k)$. The problem reduces to the estimation in each CU (i.e.: u_l) of the quantity:

$$D(k,u_l) = t_r(k,u_l) + t_0(u_l) + t_b(u_l) \quad \text{.....(5)}$$

For synchronization, $D(k,u_l)$ has to be lowered to zero for $\forall u_l \in U$ with $t_0(u_l)$ corresponding to the initial time offset measured in ms between u_l hw clock and a reference CU hw clock (i.e.: u_0) acting as a time server. $t_b(u_l)$ is equal to the booting time for acquisition processes (i.e.: Video Grabbers and Base Station Traffic Analysers) in each u_l and $t_r(k,u_l)$ is the time drift of processor mounted on the l-th CU. A NTP server (i.e.: Network Time Protocol) installed in u_0 periodically synchronizes all CUs clock. With this solution, $t_0(u_l)$ and $t_r(k,u_l)$ are set to zero by tuning an appropriate NTP broadcasting rate whereas $t_b(u_l)$ can be assumed equal to zero if fast (e.g.:>700Mhz) processors are used.

3.3 Video Object Reports Extraction

Video Objects Reports $\bar{r}^{V}_{i,m}(k)$ are evaluated by Video Metadata Extractors (i.e.: VME) at each timestamp k. VME takes as input raw video frames from synchronized grabbers; typically, chain of logical tasks can be assembled in order to process Video data [9], the first step is however a Dynamic Change Detection (see Figure 2 right) 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 (Figure.2 left). 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 tracker algorithm, that outputs to the Calibration submodule OR ($\bar{r}^{V}_{i,m}(k)$) with features:

$$\begin{aligned} \bar{f}^{i}_{1}(k) &= \bar{p}^{i}(k) = [x_I \quad y_I] \\ \bar{f}^{i}_{2}(k) &= \bar{id}^{i}(k) = [id] \\ \bar{f}^{i}_{3}(k) &= \bar{c}^{i}(k) = [c] \end{aligned} \quad \text{.....(6)}$$

where x_I, y_I are the coordinates (in pixels) of the center of mass in the *Image Plane* for the m-th object (i.e.: blob) at time k detected by i-th sensor whereas the scalars id and c respectively indicates the tracked id (progressive integer number, e.g.: 1,2..) and class of the object (integer number, e.g.: 1=human, 2=vehicle, 3=others).

3.4 Video Object Reports Spatial Alignment

Spatial alignment for Video ORs is achieved through Camera Calibration. Camera calibration [6,7] is the process by which optical and geometric features of Video Cameras can be determined. Generally, these features are addressed as intrinsic and extrinsic parameters and they allow estimation of a correspondence between coordinates in the *Image Plane* (x_I, y_I) and in the *3-D Real World Space* (x_W, y_W, z_W). After the 3-D conversion the last step is represented by the projection on 2-D *Map Plane* (x_M, y_M).

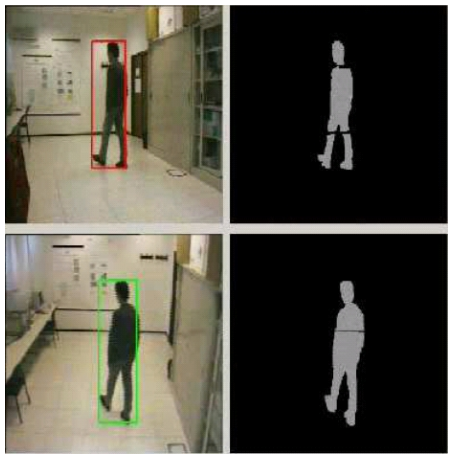

Figure 2. Right: Binary Change Detection images showing moving pixels in the current scene. *Left:* Two views for the available Video Cameras, moving object is highlighted by Bounding Boxes.

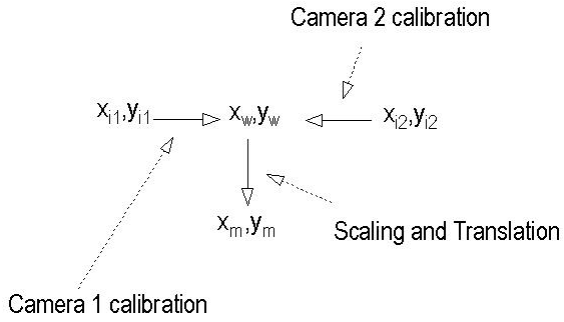

Figure 3. Joint calibration methodology.

Various method have been proposed to perform calibration: some uses non-linear optimisation techniques [8], others systems of linear equations. Camera calibration we use is based on classic Tsai method [7]. In the presented system, all video sensors have been calibrated with a common calibration strategy. In Figure 3, the chosen approach is outlined. First of all cameras are calibrated on reference images in a unique map then a common reference point has to be found in order to make the system able to switch

between the different reference systems. Origin of *World Space* (x_W^0 , y_W^0 , z_W^0) represents a good choice because it is common to the all the cameras. The alignment algorithm can be decomposed in the following steps:

1. Image Coordinates (x_I , y_I) are converted in World Coordinates (x_W , y_W , z_W) through calibration.

2. (x_W , y_W , z_W) are converted in Map coordinated (x_M , y_M) with translation and scaling transformations.

3. OR is rewritten as :

$$\overline{r}^V_{i,m}(k) = [\overline{p}^i(k), \overline{id}^i(k), \overline{c}^i(k)] = [x_M, y_M, id, c]$$

It is important to note that, in this case, id and c features which should be stationary (i.e.: identity and class of an object do not change over time) values are subject to variations over time due to induced errors in the Change Detection and Tracking steps.

3.5 Radio Object Reports Extraction:

The Radio Objects can be defined as those objects which possess electronic wireless communication facilities (e.g.: Bluetooth or WLAN cards). The wireless communication network as a part of the video surveillance system shares the information related to the objects. The Radio Objects Reports (i.e. ROR) $\overline{r}^R_{i,m}(k)$ are evaluated by receiving the signals sent by objects device to three base stations (BS). BS as a part of the wireless system are able to communicate with the radio object via receivers and they recognize RORs via their network ID. The system is based on a path-loss model of the signal power transmitted from/to APs and receivers. The observed power is converted into distances using path-loss equations, eq. (8) [11].

$$S = S_o - 10\,\alpha \log \frac{d}{d_o} \quad \text{..(8)}$$

Where S is the received power in dB, S_o is the received power at a refernce distance (*d=1 mete*r), *d* is the distance between transmitter and receiver and *α* is the path-loss exponent. Unfortunately, due to the presence of multi-path fading and noise interference in the environment, the received power is not only dependent on the path loss. Therefore, eq. (8) can be represented as eq. (9) where the observed power $(S+N)$, due to fading and path-loss is

$$(S+N) = S_o - 10\,\alpha \log \frac{d}{d_o} + X_\sigma = S + X_\sigma \quad \text{.........................(9)}$$

Where random variable X_σ represents the medium-scale fading in the channel and is typically reported to be Gaussian random variable with zero-mean (in dB) and variance σ^2, also represented as $N(0, \sigma)$ [11], [12]. The Probability Density Function (*p.d.f.*) of the received power in eq. (9) is $N(\overline{S}, \sigma)$ with $\overline{S}$ mean and Standard Deviation, *σ*.

Having estimated $(S+N)$ and X_σ, it is possible to compute the distance between transmitter and receiver using eq. (9). The distances obtained by the transceiver are tri-laterated [2] to estimate the position (x_M , y_M) in the common Map Plane and to fill an OR following the policy applied to the video case:

$$\overline{r}^R_{i,m}(k) = [\overline{p}^i(k), \overline{id}^i(k), \overline{c}^i(k)] = [x_M, y_M, id, c] \text{....(10)}$$

In this case the unstable feature is expected to be position whereas identity and class are constant over time.

3.6 Radio Object Reports Spatial Alignment

Fig 4 shows the logical functioning architecture of the WLAN network for estimation of the position $\overline{p}^i(k)$ using *RSS* features. The $RSS_{1,2,3}$ are the received signal strength of the signal at the BS sent by the object device. The Path-Loss equation (eq.9) uses this information to evaluate the distance d_{123} that is considered to be the distance between the *BS* and the object. Given the problem of presence of multi-path fading and noise in received signal, and their negative effect on position accuracy, it is desired to enhance the accuracy in two steps:

- At signal level using Pre-Post Cursor Multi-path Mitigator [4].
- At feature level using Feature Function (FF) which is created in the offline phase of the spatial alignment [13].

Spatial alignment and projection in the common 2-D map is performed in two phases:

1. *Offline phase*
2. *Online phase*

The offline phase consists of signal strength data collection at several predefined positions in the test site. Based on the $RSS_{observed}$ (i.e.: collected power strength) at known positions and $RSS_{theoeatical}$,(*i.e.* the theoreticaly computed power strength for known distances) the $RSS_{multipath+noise}$ can be computed as the difference between the above two signal values. The ratio δ , in eq. (11), is obtained for each known distance between transmitter and set of different position. The polynomial fit function, which is FF, is computed based on the collected signal measurements.
where:

$$\delta = \left(\frac{RSS_{multipath+noise}}{RSS_{observed}} \right) \quad \text{...........................(11)}$$

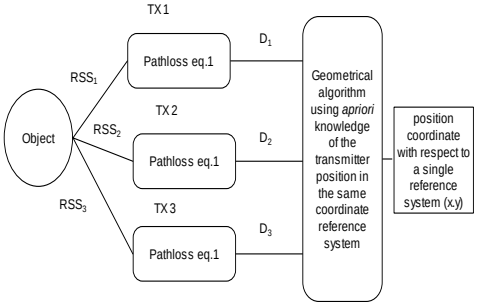

Figure 4. Logical functioning architecture of the WLAN Positioning network using RSS features.

The function FF is utilized in an online phase. During online phase, user's signal are collected by each transmitter and 2D position is computed as explained in last section. The details about position method using wlan 802.11b system can be found in [13]. In Figure 5, the view of the experimental site is shown. The BS are localized within the map and identified by dots. Three circles are centred in transmitters with radius equal to estimated transmitter distance (i.e.: d_{123}). The overlapping region of the three circles represent the most probable region where the target has to be located.

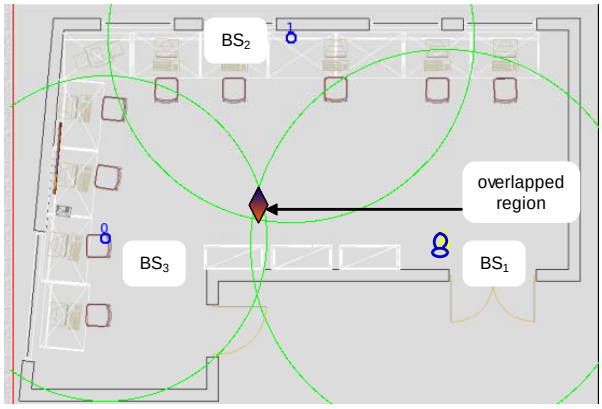

Figure 5. View of the experimental site with Base Stations and overlapping region.

4 Results

In order to validate Alignment algorithms, the following mock-up architecture has been set-up in DIBE LABs: three 802.11b WLAN Base Stations have been installed as well as two CCD-Video Cameras with 352x288 pels resolution, frame rate equal to 10 fps and partially overlapped fields of view. In addition, an actor has been asked to walk in the laboratory carrying a laptop with a WLAN card.

In Figure 6 screenshots from the output of the Video Metadata Extractor (VME in Figure 1) during the tests have been reported from the two fields of view. In figure 7, tracks coming from the VME after calibration and from Radio Metadata Extractor after trilateration are sketched in the laboratory map. As it can be seen three different tracks are available: two of them come from the cameras and they are characterized by an high accuracy in position estimation, but they not carry any identity information about the user. The third downsampled track is definitely less stable and accurate, but it contains profiling information about the tracked user associated with the MAC address of the WLAN card. Presented qualitative results are a preliminary attempt to show how Radio and Video data can be aligned in a common ground plane in order to be fused. In addition, error in localization, especially in Radio sensor is encouraging and it will allow association between the heterogeneous tracks. Statistical quantive results are expected to be collected soon on a more exhaustive number of tests; localization error will be evaluated by statistically ground thruthing the Radio Signals. Entire sequences recorded and processed during tests are available at: http://ginevra.dibe.unige.it/ISIP40/sequencesLuca.html

Figure 6. Sample processed frames indicating Blobs (i.e.: moving objects) and their centers of mass from two Fields of View.

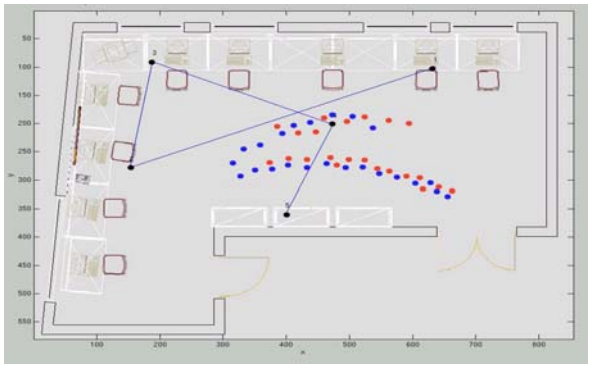

*Figure 7.*2-D map of the site in which tests took place. Three tracks are aligned and projected for Radio Objects Reports (i.e.: interconnected points) and Video Object Reports.

5 Conclusion and future work

An innovative schema towards fusion of Video and Radio Sensors mapped in a general formalism has been presented; specific algorithms for aligning data in time and space in case of Video and Radio data handling have been developed. Tests in real working conditions have been carried out showing good results.

Acknowledgements

This work has been partially co founded by Ministero dell'Istruzione dell'Università e della Ricerca within projects MIUR- PRIN, PER2 and VICOM.

References

[1] E. Waltz and J. Llinas, “Multisensor data fusion” , ISBN 0-89006-277-3, 1990 Artech House, Inc.

[2] Jeffrey Hightower and Gaetano Borriello, "Location Systems for Ubiquitous Computing," *Computer*, vol. 34, no. 8, pp. 57-66, IEEE Computer Society Press, Aug. 2001.

[3] K. Pahlavan and P. Krishnamurthy, Principles of Wireless Networks: A Unified Approach, Prentice Hall PTR, 2002.

[4] P Shan E. J. King "Cancel Multipath Interference In Spread Spectrum Communications" Wireless System Design, March 2001 Pg 49-52.

[5] Siddhartha Saha, Kamalika Chaudhuri, Dheeraj Sanghi, Pravin Bhagwat, "Location Determination of a Mobile Device Using IEEE 802.11b Access Point Signals," IEEE Wireless Communications and Networking Conference (WCNC) 2003 New Orleans, Louisiana, March 16-20, 2003.

[6] Tsai, Roger Y. (1986) “An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision,’’ Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, 1986, pp. 364–374.

[7] Tsai, Roger Y. (1987) “A Versatile CameraCalibration Technique for High-
Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses,’’ IEEE Journal of Robotics and Automation, Vol. RA–3, No. 4, August 1987, pp. 323–344.

[8] J.Renno, J. Orwell, G.A. Jones.” Towards Plug-and-Play Visual Surveillance: Learning Tracking Models”. ICIP02 (III: 453-456) September 2002 Rochester New York.

[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] Luca Marchesotti, Giuliano Scotti and Carlo Regazzoni, "Issues in multicamera dynamic metadata information extraction and interpretation for ambient intelligence", in press, Yerevan Armenia, NATO ASI 2003.

[11] G. L. Stuber, "Principle of Mobile Communication", Kluwer Academic Publishers, The Netherlands.

[12] B. Sklar, "Rayleigh Fading Channels in Mobile Digital Communication Systems Part1: Characterization," IEEE Communication Magazine, July 1997, page 90-100.

[13] R. Singh, M. Gandetto, M. Guainazzo, C.S. Regazzoni, “A Novel positioning system for static location estimation employing WLAN in indoor environment,” accepted in PIMRC, IEEE ,April. 2004.