POSITION ESTIMATION AND TRACKING WITH A NETWORK OF HETEROGENEOUS SENSORS IN VIDEO SURVEILLANCE APPLICATIONS

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

ABSTRACT

This paper presents ainnovative multi sensors system for object position estimation and tracking inspired to a model inherited from the Data Fusion domain: the Joint Directorate of Laboratories (JDL) model [1]. The problem specifically faced is location and identification of multiple users equipped with radio devices. A network of Video Sensors that cooperatively detect, segment and classify objects of interest, produce evidences regarding object location. By contrast, identity estimation is inferred using a network of Radio Sensors (IEEE 802.11 WLAN Base Stations). Results show that association is feasible between the two sets of sensors and that position estimation accuracy improves exploiting a multisensor approach.

1. INTRODUCTION

Actual trends on Surveillance Systems go into the direction of multi sensor architectures able to handle heterogeneous signals such as video, radio and audio in order to detect, localize and identify objects in an environment of interest [2]. In this paper, a complete architecture will be described according to a popular model inherited from the Data Fusion domain where Radio (i.e. WLAN 802.11b) Signals are integrated (Section 3) with Video Signals coming from multiple static CCD Cameras in a network of heterogeneous sensors. In particular, techniques to associate location information derived from radio sensors with Video sensors will be proposed (section 4). An estimation algorithm (Section 3) able to coherently fuse tracks coming from different Video Cameras, based on color, shape and dynamics information will be presented (Section 5). Results (Section 6) will show that Radio sensors are useful to preserve id of detected tracks (assuming that identity information is related to user terminal) and that are a good alternative where video sensors are not present; by contrast location estimation can be performed using Video sensors.

2. A FRAMEWORK FOR ASSEMBLING NETWORKS OF HETEROGENEOUS SENSORS

It is assumed that a set $S = \{\overline{S^c} : i = 1,...,N_S\}$ of heterogeneous sensors is divided in N_S different classes each of whom composed by a number of $N_{\overline{S^c}}$ units belonging to $\overline{S^c} = \{s_i^c : c = 1,...,N_{\overline{S^c}}\}$. Each sensor's output is processed in order to provide *Object Reports* (OR) $\bar{r}_{i,m}^{c}(k)$ for each *m-th* object present in the scene at time *k*. OR is represented as a multidimensional vector composed by different features related to the detected object:

$$\bar{r}_{i,m}^{c}(k) = [\overline{f}_1^i(k),\ldots,\overline{f}_{N_r}^i(k)]$$

with N_r the total number of features in the report. For each detected physical object tracks are instantiated and updated:

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

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

2.1 Steps towards positions extraction and tracking

The overall logic functional architecture of the proposed system is shown in Figure 1 , as it can be seen the structure is inspired to a classic model of Data Fusion systems described in [1]. Three different levels of analysis have to be performed in order to instantiate the fused tracks namely Alignment, Association and State Estimation. In order to extract salient information regarding objects of interest, a network of heterogeneous sensors is required. In the presented work the network is composed by two classes (i.e. $N_S = 2$) with S^0, S^1 respectively corresponding to CDD Video Cameras and 802.11b WLAN access points. Cardinality of the two sets is $|S^0| = 2, |S^1| = 3$, that means the sensors network is composed by 2 Static CDD Video Cameras (fixed field of view and zoom level) , and 802.11b WLAN Base Stations.

The first step in the JDL model is the time and space alignment; two modules have been inserted in order to independently pre-process information and are identified by

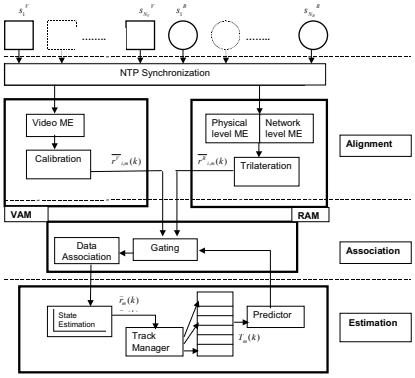


Figure 1 Logic Functional Architecture of the system composed by Data Aligment, Data Association and State Estimation

a Video Analysis Module (i.e.: VAM) that extracts metadata from a video source whereas the Radio Analysis Module (i.e.: RAM) does the same for the Base Stations. The output, namely Object Reports, is in this case addressed to as Video Object Report (VOR) and Radio Object Reports (ROR) $\bar{r}_{i,m}^{c=0,1}$; they have to be associated in relation to their features and according to general association rules.

Once different groups of ORs are evaluated by the Data Association submodule, they have to be fused in estimated reports $\hat{r}_m(k)$ by the State Estimator. Eventually, tracks $T_m(k)$ are updated or newly instantiated by the Track Manager whereas a prediction submodule feedbacks future values for ORs into the association step in order to get a closed–loop analysis.

3. NETWORK ALIGNMENT

As it can be noticed in fig.1, the first operation towards tracks generation is Data Alignment (DA); this step can be further decomposed in Temporal Alignment (TA) and Spatial Alignment (SA) in order to get a situation in which all data coming from various $s^c{}_i$ are synchronized in time and space. Whereas TA, namely temporal synchronization of all Detection Reports can be carried out with a relatively simple approach exploiting an NTP (i.e. Network Time Protocol) server [4], Spatial Alignment requires more structured and signal dependent approaches.

3.1 Detection Report extraction and spatial alignment

Given a reference coordinate space as the one depicted in Figure 3, two different strategies have been adopted to spatially align ORs:

1. Camera Calibration
2. Radio Map Calibration

The first technique turns useful to get coherent positions of VOR (Video Objects Reports: $\bar{r}_{i,m}^{c=0}(k)$) from Video Cameras at each aligned timestamp k. VAM takes as input raw Video Data from frame grabbers and it subtracts them with respect to a reference frame called "background". This process leads to the determination of moving areas (called Blob) (see Figure 2) detected in the scene, characterized by a position within the image ($\bar{p}^i(k)$) a by a "type" ($\bar{c}^i(k)$) indicating if the object is a pedestrian or a vehicle. In addition, the blob is bounded by a rectangle to which a numerical label is assigned ($\overline{id}^i(k)$) indicating the object id Figure 2. 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 and each Video Object can be characterized through Detection Report ($\bar{r}_{i,m}^{1}(k)$) taking this final form:

$$\left[\ \bar{p}^i(k)=\left[x_I \quad y_I\right],\ \overline{id}^i(k)=\left[id\right],\ \bar{c}^i(k)=\left[c\right]\ \right]$$

All steps are typically performed in the Image Plane (e.g. position is expressed in Image coordinates $x_I \quad y_I$), whereas in order to get spatially aligned reports, camera calibration strategies have to be employed to port positional information in the same reference space (i.e. map space).(Figure 3a)

By definition Camera calibration [5] 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). Camera calibration exploited in this work is based on classic Tsai method [6].

Once Video Object Reports are successfully aligned in time the same, process has to be applied to Radio Objects report (ROR).

Received Signal Strength (RSS) [4] has been chosen as representative feature for detection of Radio Objects and to implement Radio Map Calibration. RSS can be used to

determine the distance between a transmitter and a receiver by using the following path-loss equation[4]:

$$P_r = P_0 + 10\,\alpha \log \frac{d_c}{d_0}$$

Where P_r is the RRS in dBm, P_o the RSS seen at the receiver at a distance $d_0 = 1$ meter to the transmitter, d_c the distance between transmitter and c-th Base Station s^1_c and α is the path-loss exponent.

Given an emitting Object, RSS is computed at the receiver (user terminal) for all Base station and the three distances d_c (with c=0,,2) can be computed. A tri-lateration operation coherently scaled with respect to the common coordinate system (i.e. map in Figure 3) concludes the Radio Map Calibration algorithm. It takes as input the three values of d_c and outputs positional feature of i-th Radio Object assembling the following DR (Detection Report):

$$[\ \bar{p}^{i}(k) = [x_R \quad y_R],\ \bar{id}^{i}(k) = [id]\ ]$$

where *id* is the MAC address of the detected terminal.

4.HETEROGENEOUS DATA ASSOCIATION

Given aligned DRs $\bar{r}^{i}_{k,m}$, the problem of Data Association consists on the selection of the correct *track* $R^t_{K,n}$ the report belongs to. The problem can be therefore decomposed in :

1. Association of RO (Radio Object) with a given track
2. Association of VO (Video Object) with a given track

Different association metrics are used in relation to the typology of Detection Report $\bar{r}^{c=0,1}_{i,m}$. In particular the following two metrics have been found to be appropriate for the kind of treatened signals:

$$M_0(\bar{r}^{0}_{k,m}, R^t_{K,n}) = \sum_{R,G,B} \sqrt{\bar{h}^{i}_{m}(k)\bar{h}^{f}_{n}(K)}$$

$$M_1(\bar{r}^{1}_{k,m}, R^t_{K,n}) = \sqrt{\left(\bar{p}^{i,x}_{m}(k) - \bar{p}^{i,x}_{n}(K)\right)^2 + \left(\bar{p}^{i,y}_{m}(k) - \bar{p}^{i,y}_{n}(K)\right)^2}$$

The first metric evaluates the Bhattacharyya coefficient [5] between the two color histograms $\bar{h}^{i}_{m}$ $\bar{h}^{f}_{n}$ related to the m-th VO and the n-th track. A high Bhattacharyya coefficient indicates that the object is similar to the given track and that it can be associated to it. The second metric used within the scope of association is a simple Euclidean metric that is able to associate position of Radio Objects to a given track.

5. TRACKS GENERATION FROM VIDEO SENSORS

Tracks generation and updating refers to the step of state estimation that is depicted in Figure 3 (a) as the last step in the fusion schema: given set of associated DRs coming from the two cameras, a track $R^t_{K,n}$ has to be instantiated or updated. In this case, only Video Data is taken into account to estimate the position with whom updating a given track. Two different approaches are here proposed in relation to the state of occlusion (i.e.: overlapping of two or more objects in image plane) of a given Object. When the object is completely visible, a relatively simple approach has been exploited for determining the position of center of mass of the track in conditions of non-occlusion:

$$p^{i,x}_{n}(k) = \frac{1}{M}\sum_{m} p^{i,x}_{m}(k) \quad , p^{i,y}_{n}(k) = \frac{1}{M}\sum_{m} p^{i,y}_{m}(k)$$

In condition of occlusion of DRs, a more complex method [3] based on Generalized Hough Transform (GHT) and shape information is used. The GHT is a technique exploited to search arbitrary curves in an image without the need of parametric equations. A look-up table called R-table is used to model the template shape of the object. This R-table is used as a transform mechanism (Figure 2a) to build a shape model, first a reference point and several feature points (i.e.: corners point) are selected. For each feature point the orientation α of the tangential line at that point, the length r, and the orientation θ of the radial vector that joins the reference point and the feature point can be calculated. If n is the number of feature points, a 2 x n indexed table can be created using all n pairs (r, θ) and using α as index. This table is the model of the shape and it can be used with a transformation to find occurrences of the same object in other images. The high curvature points (e.g.: corners) of each blob detected in the image are extracted, and for every point, the orientation α is computed. Using α as an index for the R-table, the pair (r, θ) is extracted. Using the pair (r, θ), the possible position for the reference point can be computed and an accumulator of its positions is incremented. Although some points that do not belong to the desired shape will have similar α and they will introduce false reference points, the maximum accumulator value will occur with high probability at the actual reference point. The presented GHT variation is shown in figure 2.

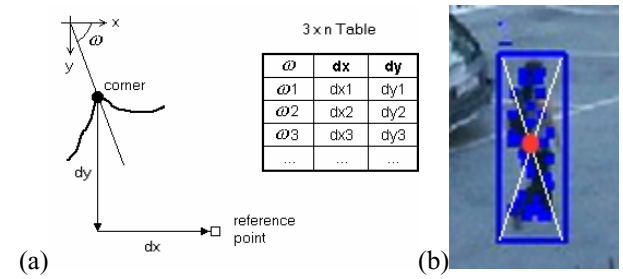


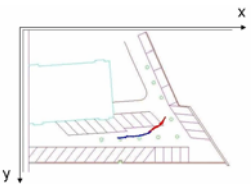
Figure 2 R-table structure (a) and result of bounding box estimation with corners (b).

Once voting spaces have been calculated, fused DR position in terms of center of mass can be evaluated by setting it to the value that received the highest number of votes in the Hough space. More details on the applied position estimation method can be found in [3].

6. RESULTS

To evaluate performances of the presented architecture the two main steps towards object position estimation and tracking have been tested separately. In particular Data Association between Objects and Tracks has been tested for Video and Radio Sensors.

Confusion Matrices (see Figure 3) have been used in order to test Data Association for Video Sensors showing a minimum spread over the diagonal and few association errors. Estimation step has been evaluated with the response of the system to situations in which multiple occluding objects are present. Key frames are reported in Figure.4 showing clean and precise tracks (trajectories) in case of evident occlusion.



	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

Figure 3 Estimation of the trajectory (a) and confusion Matrixes evaluated with neglecting speed feature (b).

Figure 4 Correct Estimation of Tracks (c) using multiple views(a-b) in condition of occlusion.

7. CONCLUSIONS

A complete framework for assembling networks of heterogeneous sensors has been presented for objects tracking and identification. It has been shown that Radio Information turns useful to reliably estimate id of a given object whereas Video Data is functional to position estimation issues. In particular proposed results confirm that the architecture improves performances of tracking in the resolution of situations of occlusions where traditional monocular approaches fail.

8.ACKNOWLEDGEMENTS

This work was performed under co-financing of the MIUR within the project FIRB-VICOM.
We would also like to thank Reetu Singh for providing support on WLAN positioning identification.

9. REFERENCES

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

[2] J. Black and T. Ellis, "Multicamera Image Tracking", 2nd IEEE workshop on Performance Evaluation of Tracking and Surveillance (PETS2001).

[3] 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.

[4] L. Marchesotti, R. Sing and C. Ragazzoni, "Extraction of Aligned Video and Radio Information for Identity and Location Estimation in Surveillance Systems" Fusion 2004 June 28 to July 1, 2004 in Stockholm, Sweden.

[5] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. Commun. Tech., COM-15:52–60, 1967.

[6] 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