A MULTICAMERA FUSION FRAMEWORK FOR MULTIPLE OCCLUDING OBJECTS TRACKING IN INTELLIGENT MONITORING AND SPORT VIEWING APPLICATIONS

Luca Marchesotti, Gianni Vernazza and Carlo Regazzoni
DIBE, University of Genoa
Via dell'Opera Pia 11a 16100 GENOVA (ITALY)
carlo@dibe.unige.it

ABSTRACT

The aim of this paper is to present a multi camera system for location estimation inspired to a model inherited from the Data Fusion domain: the Joint Directorate of Laboratories (JDL) model [1]. The problem specifically faced is the tracking of objects in two complementary applications: Intelligent Monitoring (Video Surveillance) and Sport Viewing (Football Players Tracking), where multiple occluding objects have to be successfully segmented and located using different features such as color, position and dynamics.

1. INTRODUCTION

The successful location and tracking of objects of potential interest through video processing techniques is an important step towards the understanding of events taking place in a given environment. Different issues have to be faced at architectural and algorithmic level in order to successfully detect, follow over time and classify objects such as vehicles or pedestrians.

Traditionally, such techniques have been developed within the context of mono sensor systems, which have been shown to be well performing in various applications. Anyhow, this approach is affected by different problems related to the limited system's coverage, to the unavailability of "stereo" information and mostly to the low performances in tracking non-rigid targets in conditions of occlusion. In this paper, the possibility to overcome these problems with a multisensor solution will be explored. Different reference architectures [6][9] are present in literature in the context of Intelligent Monitoring and Video Surveillance, which successfully combine multiple evidences in order to *localize* objects of potential interest. In addition, specific projects have been founded within the EU Community to port algorithms traditionally developed for surveillance purposes to the in domain of Sport Viewing to track players in football pitches [5]. Seen in this perspective, the final aim of a such systems is surprisingly close to the one of a Multiple Target Tracking application developed in the domain of Data Fusion. Anyhow, all the experiences reported in literature have been implemented within the context of Video Processing and Video Surveillance Systems without any evident link with theories [7] developed in the area of the Multisensor Data Fusion (MDF). The aim of this paper is to present a multi camera system specifically inspired to a classic DF paradigm: the Joint Directorate of Laboratories (JDL) model [1] with the precise intent of facing issues related to the overlapping of tracked objects. The paper will be structured with an overview on the formalism (section 2) used to characterize the multisensor architecture described in section 3; qualitative and quantitative results are proposed to test the system in section 4 whereas conclusions are drawn in section 5.

2. THE FORMALISM

The formalism hereinafter used assumes the presence of a set of heterogeneous sensors $S = \{s^j : j = 1,..,N_s\}$; each sensor acquires data providing *Detection Reports* (DR) $\bar{r}^j_{k,m}$ for *m–th* Object of Interest (OOI) detected at frame *k*. DR comprehend a set of features $\bar{f}^i_x(k)$ which describes at various levels of abstraction the OOI it is referred to:

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

Detection Reports related to *i-th* sensor are grouped, at time *k*, in set $R^i_k = \{\bar{r}^i_k : i = 1,....,M^i_k\}$, whereas $R_k = \{R^i_k : i = 1,...,M_k\}$ incorporates reports produced by all active sensors belonging to *S*. Tracks $R^t_{K,n}$ are instantiated and updated using Detection Reports associated to the *n-th* object present on the scene at frame *K*.

2. ARCHITECTURE LAYOUT

In fig. 1 the overall logic functional architecture of the proposed system is depicted, as it can be seen, the structure is inspired to a classic model of Data Fusion

0-7803-8554-3/04/$20.00 ©2004 IEEE.

systems described in [1]. Three different levels of analysis have to be performed towards creation of fused tracks:

1. *Report Extraction and Alignment*
2. *Data Association*
3. *State Estimation*

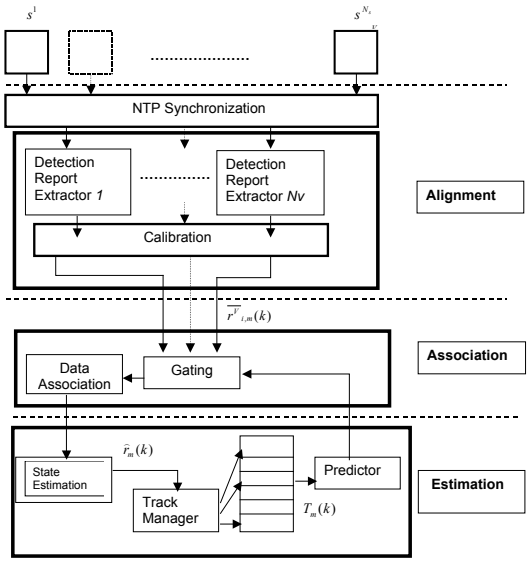


Fig.1. *Logic-Functional architecture of the proposed system.*

2.1 Report Extraction and Alignment

As it can be seen from fig.1 the first step towards fusion is the evaluation of DRs ($\bar{r}^{i}_{k,m}$) carried out by *Detection Report Extractors* and a spatio-temporal alignment. In particular, Video Frame Grabbers for A/D conversion have to be synchronized to get time aligned DRs. An NTP (i.e.: Network Time Protocol) based approach has been used for Temporal Alignment, Detection Report Extractors are allocated for each Video Camera (i.e.: sensors s^j) and they take as input synchronized images providing as output DRs with the following structure:

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

Where the various features respectively represents, position, speed, id, shape and color histogram of the detected OOI (Object of Interest). To extract these features, different logical tasks are performed. 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 (fig.2-b) to which a numerical label is assigned. In addition, center of mass of blob is evaluated and projected on a 2-D ground plane map trough a joint calibration procedure [3] that represent the spatial alignment step.

2.3 Data Association

Given an aligned DR $\bar{r}^{i}_{k,m}$, the problem of Data Association consists on the selection of the correct *track* $R^{t}_{K,n}$ the report belongs to. Relevant features for DA can be found at different levels:

- *Signal (Pixel) Level*: color histograms.
- *Object (Blob) Level*: shape, corners, position, others.
- *Event Level*: dynamics.

Different association metrics are used for association; in particular used DR's features are speed, position and color. Some features have to be preprocessed to enhance their discrimination level; in particular a linear Kalman Filter is used for position $\bar{p}^{i}_{m}(k)$, whereas a median filter is used for speed $\bar{v}^{i}_{m}(k)$. For each feature, an Association Metric M_h is build in order to give a binary output regarding the matching between the observed reports $\bar{r}^{i}_{k,m}$ and existing tracks $R^{t}_{K,n}$. Metric used to test color correspondence between observed DRs and track is represented by the evaluation of Batthacharryia coefficient [10]:

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

It measures of the distance between the two color histograms $\bar{h}^{i}_{m}$ $\bar{h}^{f}_{n}$ with respect to R, G and B channels. A high Bhattacharyya coefficient indicates that the object is similar to the given track and that it can be associated to it. Another metric used within the scope of association is a simple Euclidean metric:

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

it gives the absolute distance between center of masses of the DR and the *n-th* track $R^{t}_{K,n}$ in Image Coordinates. To take into account the dynamics of the DRs, speed vectors are used:

$$M_2(\bar{r}^{i}_{k,m}, R^{t}_{K,n}) = \sqrt{\left(\bar{v}^{i,x}_{m}(k) - \bar{v}^{i,x}_{n}(K)\right)^2 + \left(\bar{v}^{i,y}_{m}(k) - \bar{v}^{i,y}_{n}(K)\right)^2}$$

They turn out to be good features when situations as the one sketched in fig. 3a have to be faced: two pedestrians have relative small distance but colliding trajectories, therefore in this case, a position based metric will fail whereas a metric as speed can lead to the correct result. Once all metrics are evaluated, a threshold step is performed in order to get binary results out from the matching procedure:

$$o_n(M_h(\bar{r}^i_{k,m}, R^t_{K,n})) = \begin{cases} 1 & if\,(M_h(\bar{r}^i_{k,m}, R^t_{K,n})) \le th_h \\ 0 & if\,(M_h(\bar{r}^i_{k,m}, R^t_{K,n})) > th_h \end{cases}$$

Threshold step is repeated for all metrics and results are stored in a Decision Vector:

$$\bar{O}_{m_f} = [o_1, ..., o_H]$$

Association Rules are therefore applied to the Decision Vector in order to establish whether or not the DR belongs to the given track. A set A of Association Rules (AR) is therefore defined:

$$A = \{a_j : j = 0, ..., j\}$$

Commonly used ARs are MAJ (i.e.: majority) rule, AND /OR rule [7]. The final output of the Association Module takes the form of a subset of DRs, which are associated to a single track:

$$R^A_{k,m_f} = \left\{ \bar{r}^i_{k,m} \in \bar{R}_{k,m_f} : a_j(O_{m_f}) = true \right\}$$

with $R^A_{k,m_f} \subset R_k$.

2.4 State Estimation

The problem of state estimation arises when for a given set R^A_{k,m_f} of associated DRs, a track $R^t_{K,n}$ has to be instantiated or updated. In particular, features of the track have to be estimated taking into account the new associated observations (i.e. associated DRs). The principal feature to be estimated is anyhow the position of each DR, which has to be plotted in a 2-D map. Therefore, given the set of available DRs associated with the track, a relatively simple approach has been exploited for determining the position of center of mass of the track in condition 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 (i.e.: overlapping of two or more objects) of DRs, a more complex method [4] 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 (fig. 2-a). To build the R-Table, first a reference point and several feature points (i.e.: corners point) of the shape 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. Using a voting technique localizes the shape. 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 position is incremented. Although some points that do not belong to the desired shape will have similar α and 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.

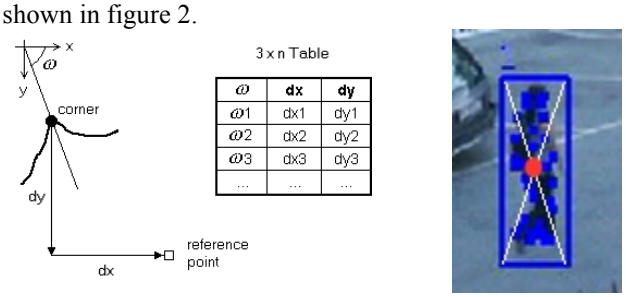


Figure 2. R-table structure (a) and result of bounding box estimation with corners (b).

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 [4].

3. RESULTS

To evaluate performances of the presented architecture, qualitative and quantitative results are proposed. Confusion Matrixes have been used to test Data Association in two different situations: the first one (Table 1-(a)) reproduces the case in which all features vectors in Detection Report are taken into account (i.e.: position, distance and speed) and 4 different tracks (i.e.: objects) 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 between detected objects and tracks whereas, if speed feature is neglected in association (Table 1-(b)), spreading outside the diagonal is evident and association is corrupted. Estimation step has been benchmarked evaluating the response of the system to situations in which multiple occluding objects are present. Key frames are reported in fig.2 showing clean and precise tracks (trajectories) in case of evident occlusion. Fig. 3 shows that the system can be successfully applied in domains alternative to Monitoring.

	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. Confusion Matrixes evaluated with all features in DR (a) and neglecting speed feature (b).

The reported frames are taken from a football sequence which is entirely available: http://ginevra.dibe.unige.it/ISIP/results.html. It shows multiple occlusions simultaneously engaged and tracks which are correctly instantiated and updated.

4. CONCLUSIONS

A complete framework for Multicamera Tracking has been presented exploiting different features such as color, position and speed of detected objects; the innovative approach derived from the Data Fusion domain has been shown to be well performing in different applications. In particular proposed results confirm that the architecture improves performances of tracking in the resolution of situations of occlusions where traditional monocular systems fails.

5. REFERENCES

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

[2] 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.

[3] 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

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

[5] http://inmove.erve.vtt.fi/

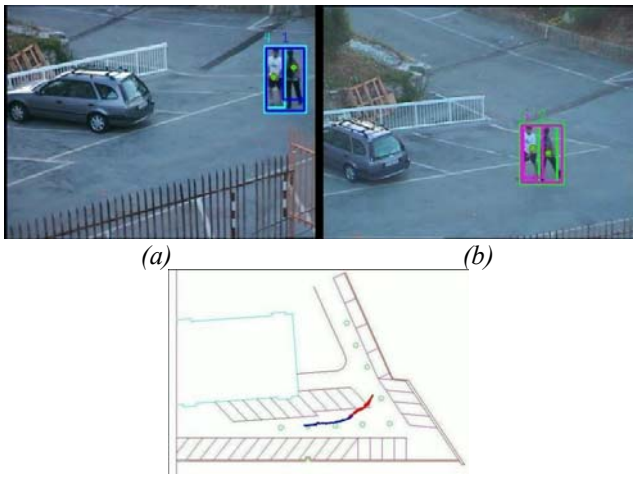
[6] J. Black and T. Ellis, "Multicamera Image Tracking", 2nd IEEE workshop on Performance Evaluation of Tracking and Surveillance (PETS2001).

[7] P. K. Varshney, "Distributed Detection and Data Fusion," N.Y.: Springer-Verlag, 1997.

[8] L.Marcenaro, F.Oberti, C.S.Regazzoni, "Multiple objects color-based tracking using Multiple-cameras in complex time-varying outdoor scenes", 2nd IEEE Int. Workshop on Performance evaluation of tracking and surveillance, Kauai, Hawaii, USA, Dec. 2001. (2001 IEEE).

[9] B. Collins, Lipton, Fujioshi Kanade "A system for video surveillance and monitoring", Proc. IEEE Vol 89 1456-1477 oct 2001

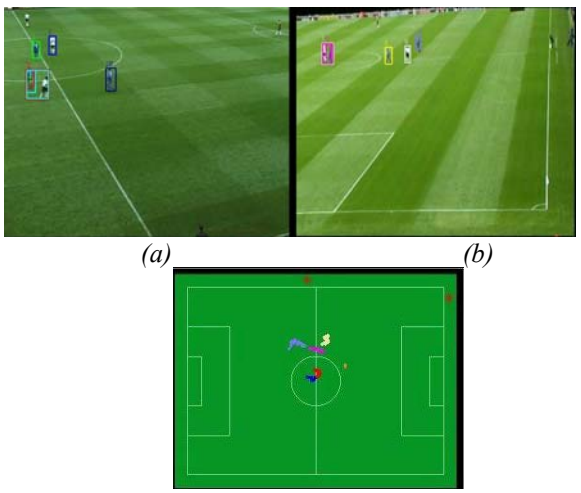
[10] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Trans. Commun. Tech*., COM-15:52–60, 1967.



(a) *(b)*

(c)

Fig.3: Correct Estimation of Tracks (c) using multiple views(a-b) in condition of occlusion.



(a) *(b)*

(c)

Fig.4: Football Sequences tracking (c) using multiple views (a-b) in condition of occlusions.