A Multi-Feature Object Association Framework for Overlapped Field of View Multi-Camera Video Surveillance Systems

Stefano Piva, Alessandro Calbi, Daniele Angiati and Carlo S. Regazzoni
D.I.B.E. – University of Genoa
piva@dibe.unige.it

Abstract

This work describes a data fusion technique to improve performances in objects localization and tracking for automatic video surveillance systems.

The developed strategy is designed to perform well in case of interaction among objects, i.e. when the moving objects to track, and whose position we want to locate on the common map reference system, result superimposed in the image plane. In order to solve such complex situations, different kind of techniques have been integrated but the focus of the paper is on the data association step in the fusion chain. As discussed in the text, failing in the association phase means computing wrong position during fusion process.

The performances of the developed technique has been evaluated on sequences of real images and experimental results show the validity of the approach in the reduction of association errors during occlusion phases.

1. Introduction

Recent years events have dramatically emphasized security issues in critical areas. One of the consequences of these events is a boost in safety measures demand and therefore in video-surveillance systems research. In the same period, image processing techniques have taken advantage from the increased computing capabilities and the lowering of hardware costs. The result is a fervent scientific community exploring all the processing levels and many new solutions to make automatic video surveillance reliable and feasible as a (partial) substitute of human control[1]. In this field, one of the most exploited research line nowadays is the use of redundant information provided by the use of sensors' duplication: in order to assure the proper coverage upon wide areas and thanks to the availability of fast, cheap and standard wired or wireless digital transmission solutions, many multi-camera systems have been designed and many tested to find different effective ways to exploit the provided additional information. Nevertheless sensors duplication is not only used to widen the coverage area but also to increase image processing performance and precision. In this work we deal with this last objective: to improve the system's ability to locate interesting moving objects' position on a calibrated map, through the joint use of a set of overlapped field of view cameras [2]. Treating tracking and positioning issues, we rely on the use of two sensors with possibly different dynamic occlusions features to better estimate the desired objects' positions.

Each single camera processing chain is based on change detection techniques, computed over an updated static background. The objects are then labeled and tracked along time performing relative recognition.

2. Motivation

In this paper we thus deal with the problems connected to the joint use of several sensor and the way to achieve better performance through information duplication and not to fail accomplishing one of the data fusion objectives: to obtain a result not worse than the one achievable with a single sensor [3].

Specifically the influence of the *Data Association* phase in the fusion procedure is addressed, starting from the considerations about the importance it has on the entire process: trying to achieve better results through the use of redundant information coming from two different sensors is certainly harmful if these information come from wrongly associated objects. Common systems often fail objects' association when tracked targets are not well separated in the image plane (*occlusion phases*).

0-7803-9385-6/05/$20.00 ©2005 IEEE

3. Multi-Camera Data Fusion

In order to exploit the redundant information obtained by the use of multiple cameras then we refer to a classical data fusion approach [4][5] structured in three fundamental steps:

- *Data Alignment*
- *Data Association*
- *State Estimation*

In the next three paragraphs data alignment is briefly discussed to introduce to the problem of data association, main topic of the present work. The state estimation strategy of the described system is then introduced.

3.1 Data Alignment

Data alignment is needed in order to make the data comparable: dealing with video cameras, this step issues are related to temporal and spatial alignment.

3.1.1 Temporal alignment

All the data flows coming from each sensor have to be synchronized to compare features referring to the same instant. This can be obtained thanks to a NTP (Network Time Protocol) [6] server. Using different kinds of hardware, cameras and in case frame grabbers, it is needed to set the system in order to have the same output frame rate.

3.1.2 Spatial alignment

Spatial alignment is obtained thanks to the joint cameras calibration procedure. Calibration consist in the determination of the correspondences among the image planes and the absolute "world coordinates", exploiting geometric and optic features of each sensor; this means it consist in the attaining of the relationships connecting the "real" locations of objects in the scene and their position in the map reference system. The most known calibration technique and the one also used in the development of the presented system, is described in [7].

3.2 Data Association

The core of the proposed system is represented by the multi-camera association strategy.

Data association is defined as an m-ary decision process among the objects in the fields of view of the used cameras.

The idea is to exploit a number of different features to let the system autonomously adapt the association to different situation occurring in the scene. The common way to manage association in different conditions for example of movement and lightning, is to choose different features if the objects to track are moving or if they are well separated in the image plane. An help also often comes from the use of time-based regularization data filters.

Despite of this approach, we aim at optimizing the sole association in each single frame considered for itself. Obtaining this means to allow an even better performance when we subsequently apply memory-based differential techniques.

The use of different features has the advantage to extract in every instant and among the others the better discriminating feature, which will be responsible of the greater separation among the classes we are trying to distinguish in the decision process.

In the current system implementation we make use of feature functions based on

- Position
- Speed
- Shape factor
- Chromatic characteristic

The first two functions are then measured in the map reference system common to all the interested sensors, while shape and color are features proper of the image plane.

In order to be able to manage such different data and obtain a coherent representation, we define independent similarity functions connected to these four pieces of information obtaining from each one of them an autonomous similarity coefficient yielding continuous values distributed between 0 and 1.

In this way two different objects can -as instance- obtain a maximum similitude (i.e. 1) coefficient for what concerns the sole speed vector direction factor if they both are still in the scene.

This apparent problem is solved by the contribution of the other features participating to the computation of global similarity coefficient: the attempt to find the correct association for an object belonging to a first camera field of view with two objects belonging to a second sensor will not be compromised when the mutual feature computation provides the same influence to the total evaluation. This because the decision classes separation will be determined by other features resulting more reliable in that particular situation.

The following sections are dedicated to the description of the similarity functions defined for the four listed objects' properties.

3.2.1 Position Similarity Function *f(P)*

Position is the mostly used feature to associate objects when the system is able at representing data coming from different cameras in the same reference system –an area map- through calibration techniques. Experiments demonstrate that without regularizing

position through time history properties, the obtained data in classical automatic video-surveillance application environments are not stable. Moreover when the monitored objects are interested by an occlusion situation in the image plane, this feature becomes unreliable and its exclusive use does not provide correct results. As previously stated, in the presented system, position is important to associate different instances of the same object but it is only one of the four features cooperating to obtain the final similarity factor which will be mostly influenced by the results of other functions when position fails.

Considering two objects observed by two of the different cameras (indexes i and j) composing the system sensor set and whose coordinates are reported in the common reference system represented by the area map, $o_m = (x_m^i - y_m^j)$ and $o_n = (x_n^i - y_n^j)$, we simply compute their Euclidean distance

$$d_{m,n} = \sqrt{(x_m^i - x_n^j)^2 + (y_m^i - y_n^j)^2} \quad (1)$$

and use this distance in the following exponential function

$$f_{m,n}(P) = e^{-\alpha d_{m,n}} \quad (2)$$

with $\alpha \geq 1$ *position difference amplification factor*, to obtain a position similarity coefficient giving 1 as maximum resemblance value when $d = 0$, to quickly reduce its entity in dependence of α.

3.2.2 Velocity Function *f(V)*

The second map reference system-based feature introduced is related to direction of the velocity vector. This is in many cases a very discriminating attribute, providing good results when two objects approach and their detected blobs fuse in a single difference area making the position feature useless: as anyone of the used features, speed vector direction is particularly useful in specific situation and fails in others. This is the reason to introduce multi-feature decision.

To obtain a motion direction similarity factor we calculate the vertical and horizontal components of the speed in the map plane and compute a normalized scalar product to achieve the difference angle cosine and then translate it and divide by two in order to attain a coefficient distributed between 0 and 1, as desired.

With notations similar to the position function, we indicate with $\mathbf{v}^i_{\mathbf{x},m}$ the horizontal component of the *m* object in the image plane of the *i-th* video sensor:

$$f_{m,n}(V) = (1 + \frac{v^i_{x,m} v^j_{x,n} + v^i_{y,m} v^j_{y,n}}{v^i_m v^j_n}) / 2 \quad (3)$$

$f(V)$ is equal to zero when the interested objects move in opposite direction and to one when they go through parallel paths. To regularize the results, objects speed must exceed a threshold to be taken into account. This is needed to avoid considering small shifts in the projection of the position due to low level image processing tasks errors in the detection of the center of mass as actual movements.

3.2.3 Shape Factor Function *f(S)*

The third feature concurring to the determination of the similarity decision parameter is based on a simple property of the video surveillance application typical objects silhouettes: the bounding box containing the shape of a vehicle as a car has different proportions than the ones of a pedestrian. This property, though not particularly discriminating on its own, is useful to contribute to the validation of correspondence hypotheses. The main benefit of the use of the ratio between height and width of the bounding box is, along with its extreme simplicity, its invariance with the distance and, consequently with the dimension of the shape.

To make use of this feature in the definition of the unique similarity factor we consider the bounding box's diagonal angle and compare the candidate couples (figure 1) again through the use of the normalized scalar product.

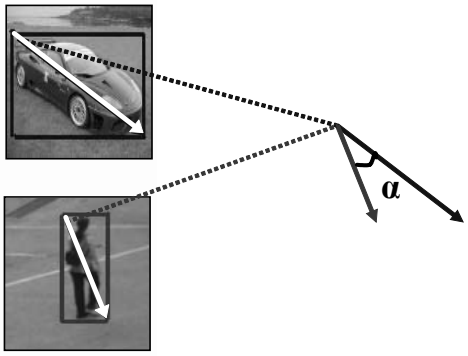

Figure 1: Shape proportions comparison through normalized scalar product of the bounding box's diagonal angle

If $\mathbf{d}^i_{\mathbf{x_m}}$ is the dimension along the X axis of the *m-th* object in the *i* camera image plane,

$$f_{m,n}(S) = \frac{d^i{}_{x_m} d^j{}_{x_n} + d^i{}_{y_m} d^j{}_{y_m}}{d^i{}_m d^j{}_n} \quad (4)$$

is the expression of the normalized scalar product providing one as result of the comparison of two shapes presenting the same proportions and zero as asymptotic value deriving from the geometric aberration of an infinitely high bounding box compared with an infinitely wide other shape.

3.2.4 Chromatic Similarity *f(C)*

The last feature here exploited to speculate on the similarity of two association candidates is the widely employed color similarity factor. This factor is expressed through the use of the Bhattacharya coefficient to correlate the color histograms of moving objects in different camera image planes.

$$f_{m,n}(C) = \frac{\sum_{c=R,G,B} \sqrt{h^i{}_c(m) \cdot h^j{}_c(n)}}{3} \quad (5)$$

To optimize the histogram computation, experiments have been conducted to set the main parameters in accordance with the application: in particular tests demonstrated that the use of more than 32 bins in the computation of the color histogram in outdoor video surveillance sequences does not yield significant benefits in spite of increased computational load requirements. Bhattacharya comparative coefficient as well as the other previously described similarity factors, provides results in a continuous scale between 0 (completely different histograms) and 1 (identical histograms), allowing the direct comparison and joint use we seek.

3.2.5 Object Similarity Coefficient *(OSC)*

Once the features to use in the association step have been chosen and the single-feature similarity functions have been defined in order to attain the common function image $\in [0;1]$, it is very easy to put them together in a single probability: a simple mean value computed on the *N* selected functions can be used to provide the relative similarity probability between two objects *m* and *n* in the field of view of video sensors *i* and *j*

$$OSC_{m,n} = \frac{\sum_{l=1}^{N} f^l{}_{m,n}(.)}{N} \quad (6)$$

For our chosen feature set, it becomes:

$$OSC_{m,n} = \frac{f_{m,n}(P) + f_{m,n}(V) + f_{m,n}(S) + f_{m,n}(C)}{4} \quad (7)$$

but the setting up of the framework allows very easy introduction of other objects' characteristics in the decision comparison.

To apply the criterion and choose the correct associations we seek for the highest values for each object in a camera field of view compared to all the objects in all the cameras image planes with a field of view overlapped with the first. A graphical representation of a simple example is shown in figure 2: three people are seen in the common field of view area (*C* in the map image) of camera *i* and camera *j*; OSC value is calculated for all the possible associations among the objects and the maximum result is selected for each object. In the example of figure 2, the evaluation is depicted for object *m* and the winning association $\mathcal{A}(m,n)$ is highlighted.

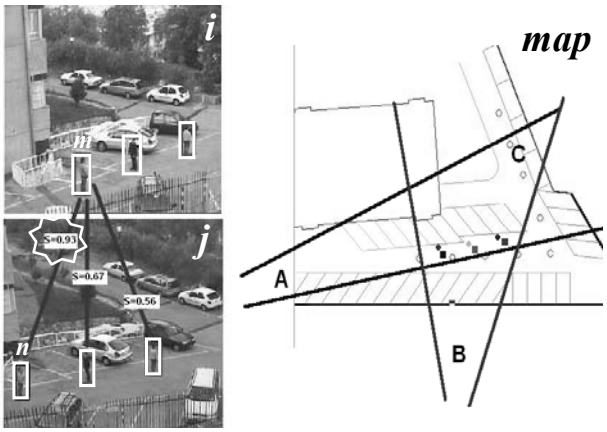

Figure 2: Example of the OSC value comparison for object *m* in camera *i* field of view with objects belonging to the overlapped monitored area *C* of camera *j*

For an object *m* the associated object is therefore decided on the bases of

$$\mathcal{A}(m,i) = \max_i (OSC_{m,i}) \quad (8)$$

In candidates selection for multiple objects association is not always easy to avoid the likely conflicts. To resolve the ambiguities we operate on similarity matrices where many strategies can be implemented for example to force the association even in presence of very "weak" *OSC* values or, on the contrary, to avoid associating objects in critical

applications, yielding a *NC* (*Not Classified*) response. The problem is shown in the simple example of Table 1 (2 cameras, 3 objects): the bold values present an ambiguity because if either we look for maxima in horizontal search direction or in vertical we find different results.

Among the several choices in the current implementation of our system we prefer not to force weak associations, using a threshold to discard couples presenting OSC values below 0,5.

	1*i*	**2*i***	**3*i***
1*j*	0.61	0.75	0.80
2*j*	0.42	***0.85***	0.37
3*j*	0.25	***0.90***	0.63

Table 1: Example of similarity matrix with highlighting of an ambiguity situation

3.3 State Estimation

Once data are aligned and objects associated, the State Estimation phase performs the actual redundant information exploitation: when the single cameras positioning data are available, they can be fused simply through the use of mean values. But when objects are not well separated in the image plane, a little more care must be put in the estimation phase.

In our 2-camera system we consider 3 cases:

- if the objects to associate are well separated in both the fields of view, we use the position mean value
- if the objects result occluded in the field of view of one of the sensors, we use the position computed by the other
- if both the fields of view present occlusions, we apply the location data related to the objects' couple with the "strongest" *OSC* value in the association phase.

4. Results

To evaluate the association rule performances we tested several coupled sensors outdoor video sequences containing variable numbers of objects.

In this section we present the most significant results related to sequences with 2 and 4 moving objects acting and sometimes mutually occluding, in the form of association rate confusion matrices: in the principal diagonal cells the rate of correct association is reported while the crossing values in the other cells define the wrong associations. Some associations were discarded defining the data belonging to the NC class.

Table 2 contains the result matrix for the 2 objects sequences using the sole position feature (as in the old system implementations) to provide a comparison. Table 3 demonstrates the improvements on the same sequences using the described 4 features OSC based system.

	1i	***2i***	***NC***
1j	91.8	5.2	3.0
2j	6.2	88.9	4.9
NC	2.0	5.9	

Table 2: Association rate confusion matrix: 2 cameras, 2 objects sequences,
$OSC_{m,n} \equiv f_{m,n}(P)$

	1i	***2i***	***NC***
1j	98.2	0.0	1.8
2j	1.0	99.0	0.0
NC	0.8	1.0	

Table 3: Association rate confusion matrix: 2 cameras, 2 objects sequences,
$OSC_{m,n} = f(P,V,S,C)$

The same comparison is then presented in tables 4 and 5 again respectively reporting results with the use of the sole position feature and with the complete set of the chosen 4 features.

	1i	***2i***	***3i***	***4i***	***NC***
1j	90.0	3.3	0.0	0.0	6.7
2j	3.3	87.7	0.0	0.0	9.0
3j	0.0	0.0	85.5	4.5	10.0
4j	0.0	0.0	3.3	91.1	5.6
NC	6.7	9.0	11.2	4.5	

Table 4: Association rate confusion matrix: 2 cameras, 4 objects sequences,
$OSC_{m,n} \equiv f_{m,n}(P)$

Last presented association results are referred to a sample of 4 objects sequence observed along time: ninety frames contain two occlusion phases where association errors are much more frequent due to the failing of the position and often of the color features. The Y axis has discrete values $\in \left[0, \frac{1}{4}, \frac{2}{4}, \frac{4}{4}\right]$

with the meaning of the number of erroneous associations on the total of four.

	1i	*2i*	*3i*	*4i*	*NC*
1j	97.7	0.0	0.0	0.0	2.3
2j	0.0	97.7	0.0	0.0	9.0
3j	0.0	0.0	93.3	1.1	5.6
4j	0.0	1.1	0.0	98.9	0.0
NC	2.3	1.2	6.7	0.0	

Table 5: Association rate confusion matrix: 2 cameras, 4 objects sequences,
$OSC_{m,n} = f(P,V,S,C)$

It is easy to be convinced of the higher number of errors presented in figure 3 where the sole position is used, comparing to figure 4, where the 4-feature set is exploited.

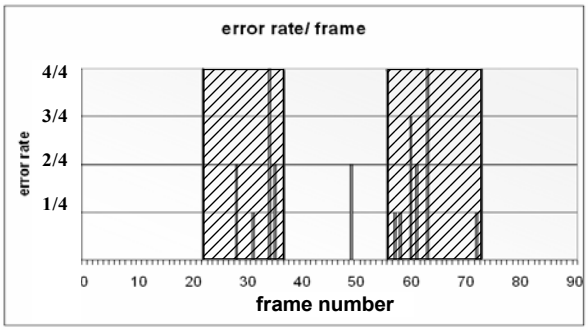

Figure 3: Example sequence containing two occlusion phases (shaded areas);
$OSC_{m,n} \equiv f_{m,n}(P)$ **; histogram lines indicate the instantaneous number of erroneous associations**

5. Conclusions

In this paper the issue of different objects instances association in different sensors is addressed. The problem rises in multi-sensor video surveillance applications especially in cases of mutually occlusion by the objects. In calibrated systems this leads to errors in the position detection on the map common reference system.

The idea is to exploit different independent similarity functions with the characteristic of having image in [0; 1]. In this way results are comparable and a global similarity value (*Object Similarity Coefficient*) for each couple is easily defined as the mean value of the computed single-feature factors.

The system performance are evaluated without the help of memories and state filtering through time-related information: demonstrated the good performance of the technique in independent frame-by-frame working situation, the addition of filtering and data time regularization will further improve results.

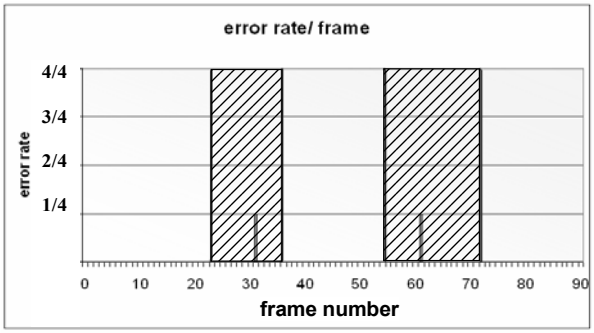

Figure 4: Example sequence containing two occlusion phases (shaded areas);
$OSC_{m,n} = f(P,V,S,C)$ **; histogram lines indicate the instantaneous number of erroneous associations**

6. Acknowledgments

This work was performed under co-financing of the MIUR within the project FIRB-VICOM.

7. References

[1] C. S. Regazzoni, R. Visvanathan, and G. L. Foresti, "Scanning the issue technology - Special Issue on Video Communications, processing and understanding for third generation surveillance systems," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1355–1367, October 2001.

[2] R. T. Collins, A. J. Lipton, and T. Kanade, "Introduction to the special section on video surveillance," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 745–746, August 2000.

[3] D. L. Hall and A. K. Garga, "Pitfalls in data fusion (and how to avoid them)," in *Proceedings of Fusion'99*, July 1999.

[4] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion,"*Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, January 1997.

[5] M. E. Liggins, C. Chong, I. Kadar, M. G. Alford, V. Vannicola, and S. Thomopoulos, "Distributed fusion architectures and algorithms for target tracking," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 95–107, January 1997.

[6] "NTP: The Network Time Protocol", [Online] available at http://www.ntp.org/

[7] R.Y. Tsai, "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, Vol. RA-3, No. 4, August 1987, pages 323-344.