Quality-Based Fusion of Multiple Video Sensors for Video Surveillance

Lauro Snidaro, Ruixin Niu, Gian Luca Foresti, *Senior Member, IEEE*, and Pramod K. Varshney, *Fellow, IEEE*

***Abstract*—In this correspondence, we address the problem of fusing data for object tracking for video surveillance. The fusion process is dynamically regulated to take into account the performance of the sensors in detecting and tracking the targets. This is performed through a function that adjusts the measurement error covariance associated with the position information of each target according to the quality of its segmentation. In this manner, localization errors due to incorrect segmentation of the blobs are reduced thus improving tracking accuracy. Experimental results on video sequences of outdoor environments show the effectiveness of the proposed approach.**

***Index Terms*—Data fusion, object tracking, segmentation quality, video surveillance.**

I. Introduction

Interest in automatic surveillance systems is rapidly gaining momentum over the last few years. This is due to an increasing need for assisting and extending the capabilities of human operators in remotely monitored large and complex spaces such as public areas, airports, railway stations, parking lots, bridges, tunnels, etc. The last generation of surveillance systems was designed to cover larger and larger areas dealing with multiple video streams from heterogeneous sensors [1], [2].

The ultimate goal of these systems is to automatically assess the ongoing activities in the monitored environment flagging and presenting to the operator suspicious events as they happen in order to prevent dangerous situations. A key step that can help in carrying out this task is analyzing the trajectories of the objects in the scene and comparing them against known patterns. In fact, the system can be trained by the operator with models of normal and suspicious trajectories in the domain at hand. As recent research shows, this process can even be carried out semiautomatically [3].

Therefore, a successful video security application requires an underlying robust and accurate tracking of the objects in the scene. That is, at each time instant the system needs to recursively estimate and predict the objects' state, including their positions and velocities, based on sensors' measurements that arrive sequentially. This problem is usually solved by the well-known Kalman filter for a linear system with Gaussian noise, and by the extended Kalman filter or particle filter for a nonlinear system. To implement any of these algorithms, two kinds of noises should be modeled appropriately: process noise and measurement noise. Predicting the next state of the target also requires a model of the targets' motion. However, this cannot be modeled exactly since sometimes targets are deliberately noncooperative and maneuver in an unpredictable manner, as encountered in military and surveillance applications. Therefore, in target tracking applications, the state process noise is employed to model the uncertainty of the motion of the objects. As a result, the actual measurements can differ substantially from the predictions made by the tracking filter based on a predefined dynamic motion model. In a video surveillance system, measurement noise is primarily due to the acquisition process and by reference plane transformations. This noise can severely degrade the accuracy of a target's current state estimate, therefore also affecting the prediction phase that follows, eventually yielding a coarse estimation of the target's trajectory.

The uncertainty of the target motion model can be reduced by adopting multimodel filtering techniques like the interacting multiple model (IMM) estimator [4]. The IMM estimator is a very successful tracking scheme particularly for tracking targets that maneuver from time to time. Measurement noise may be reduced by either adopting more accurate sensors or by placing more sensors. The latter case involves the use of data fusion techniques [5] to improve estimation performance and system robustness by exploiting the redundancy provided by multiple sensors observing the same scene. With recent advances in cameras and processing technology, data fusion is increasingly being considered for video-based systems. In addition, the main hurdle of the additional computational requirements has been removed by the tremendous processing power of today's CPUs. Intelligent sensors, which are equipped with microprocessors to perform distributed data processing and computation, are also available and can reduce the computational burden of a central processing node.

The problem of tracking humans and vehicles with multiple sensors has already been investigated [2]. However, the reliability of the sensors is never explicitly considered. In a video surveillance system that employs multiple sensors, the problem of selecting the most appropriate sensor or set of sensors to perform a certain task often arises. The task could be target tracking, tape recording of a suspicious event, or triggering of an alarm. It would be desirable to have a system that could automatically pick the right camera or set of cameras. Furthermore, if data from multiple sensors are available and data fusion is to be performed, results could be seriously affected in case of a malfunctioning sensor. Therefore, a means to evaluate the performance of the sensors and to weight their contribution in the fusion process is definitely required.

An adaptive multicue multicamera fusion framework based on democratic integration [6] is presented in [7]. Fusion is performed by taking into account sensor reliability but there is no direct sensor quality assessment. Instead, the realiability of a source is estimated by measuring the distance between each source's estimate and the fused estimate, which is determined by the sources' estimates. This is based on the assumption that the majority of sensors are producing reliable estimates, which cannot always be taken for granted.

The major contributions of this correspondence are the following: 1) the employment of multiple video sensors to enhance target localization accuracy through data fusion; 2) the development of a new quality function to dynamically assess the performance of the sensors for each target; 3) explicit consideration of the accuracy of the sensors in the fusion process through a weight function. In addition, we further develop the confidence metric introduced in [8], by taking into account also blob connectivity as a quality factor.

II. Sensor Evaluation and Weighting

If multiple sensors provide redundant information on a target's state then data fusion may be exploited to improve tracking accuracy. The

Manuscript received April 13, 2006; revised September 26, 2006. This work was supported in part by the Italian Ministry of University and Scientific Research within the Project PRIN06 (Ambient Intelligence: Event analysis, sensor reconfiguration, and multimodal interfaces), in part by the European Project SEC6-SA-204400 (HAMLeT), and in part by the Army Research Office (ARO) under Grant W911NF-06-1-0250. This paper was recommended by Associate Editor Q. Zhu.

L. Snidaro and G. L. Foresti are with the Department of Mathematics and Computer Science, University of Udine, 33100 Udine, Italy.

R. Niu and P. K. Varshney are with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2007.895331

fusion process calls for a weighting policy to be adopted since not making any distinction between the data provided by the different sensors could lead to filter instability and coarse estimates, particularly in presence of poorly performing or faulty sensors [5].

Sensors may have different internal characteristics so that their ability to detect a target could be differently affected by distance and illumination conditions. There is a lot of research being carried out in the field of objective image quality evaluation. A review of some of the most effective metrics developed so far can be found in [9]. However, these metrics are generally used to estimate the quality of degraded images due to noise or compression when the flawless original image is available. Assuming that a reference segmentation is available, these metrics are claimed to be applicable to vision systems, such as surveillance ones, to evaluate the quality of segmentation [10], [11]. This is called "objective relative evaluation" [11] and is clearly not the case for most of the real-world systems.

The quality function described here does not rely on flawless reference images and is meant to evaluate only what is really available from the system, the video signals. The proposed function gives a score to every blob detected by each sensor by processing the difference between the current image and the one maintained as reference one (background). This function marks every blob with a score that gives a value to the segmentation of the detected object. Since segmentation errors can eventually translate into localization errors, the position measurements of a given target obtained from blobs with low values of the quality function are considered less reliable.

A. Quality Function

The quality function ϕ presented here represents an improvement over the appearance-ratio function described in [8], since: 1) it also considers the connectivity of the blob, and 2) a more lenient behavior has been attained by modifying the normalization factor.

The following notation will be used in the equations:

$\mathbf{A}$	Multichannel color image $\mathbf{A}$.
$\mathbf{A}_c$	Channel c of multichannel image $\mathbf{A}$.
$A(x,y)$	Color pixel at position (x,y) of $\mathbf{A}$.
$A_c(x,y)$	Component c of pixel (x,y) in $\mathbf{A}$.

The function ϕ gives a value to the degree of confidence associated with $\mathbf{b}^s_{j,k}$, that is the jth blob extracted at time kT from the sensor s, where $T = (1/25)$ s is the sampling interval of video cameras

$$\phi\left(\mathbf{b}^s_{j,k}\right) = w_1 \upsilon\left(\mathbf{b}^s_{j,k}\right) + w_2 \chi\left(\mathbf{b}^s_{j,k}\right). \tag{1}$$

The measure is a combination of the two functions υ and χ with associated weight parameters w_1 and w_2. During the experiments, no preference was expressed thus reducing (1) to the straight arithmetic average of the two measures by choosing $w_1 = w_2 = 1/2$. The first component models how much the blob is discernible from the background and is defined as follows:

$$\upsilon\left(\mathbf{b}^s_{j,k}\right) = \frac{\sum_c \sum_{x,y\in\mathbf{b}^s_{j,k}} D_c(x,y)}{\sum_c \sum_{x,y\in\mathbf{b}^s_{j,k}} \epsilon\left(B_c(x,y)\right)} \tag{2}$$

where $\mathbf{D}$ is the difference map obtained as absolute difference between the current image $\mathbf{I}$ and the reference one $\mathbf{B}$. The numerator is therefore a sum over the values of the pixels in the difference image belonging to the detected blob. The denominator represents a sum over the possible spread of the difference between $B(x,y)$ and $I(x,y)$ for the pixels belonging to $\mathbf{b}^s_{j,k}$. For both the numerator and denominator, the sums are carried out spatially over the pixel coordinates and then chromatically over the number c of the color bands of the involved images. In this correspondence, three-band red, green, and blue color images are employed. The function ϵ calculates the possible spread for each pixel in the following way:

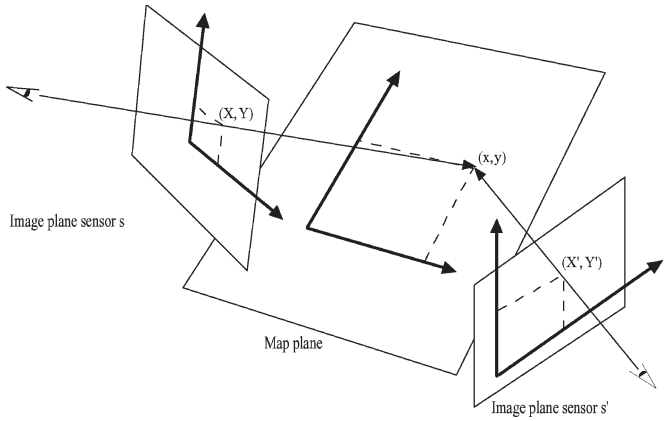

Fig. 1. Homographic transformation. Image planes of sensors s and s' are mapped to a common reference frame.

$$\epsilon\left(\mathbf{B}_c(x,y)\right) = \max\left(B_c(x,y), 255 - B_c(x,y)\right). \tag{3}$$

For example, $B_1(x',y') = 75$ means that the value of the first color component (red) of the pixel at position (x',y') of the reference (background) image is 75. The maximum difference that the current image in the red color component of pixel (x',y') can display in this case is $255 - 75 = 180$ dimmer color tones. Therefore, ϵ computes the maximum difference (in brighter or darker tones) that a pixel (x,y) in the current image could display compared to the corresponding pixel in the background.

The function χ measures the goodness of the segmentation of a blob through its connectivity. A blob is supposed to have been correctly extracted if it is composed of a single connected component. Therefore, given the number $n(\mathbf{b}^s_{j,t}) \geq 1$ of connected components of blob $\mathbf{b}^s_{j,t}$, the function χ yields the value 1 for $n(\mathbf{b}^s_{j,t}) = 1$, and decreases, in a Gaussian manner, for increasing $n(\mathbf{b}^s_{j,t})$

$$\chi\left(\mathbf{b}^s_{j,k}\right) = e^{\frac{-\left[n\left(\mathbf{b}^s_{j,k}\right)-1\right]^2}{2\sigma^2}}. \tag{4}$$

During the experiments, the standard deviation 0 takes values in the interval $4 \leq \sigma \leq 6$, depending on the distance of the targets from the camera, computed taking into account the projection on the ground plane (Fig. 1). The homographic transformations between the camera plane and the ground plane are determined during the initial setup of the system. The operator establishes the correspondences between salient points on the image planes and the ground plane. The homographic transforms are then easily found [12]. Closer targets are more likely to exhibit cracks in their segmentation. This is due to the fact that closer targets cover a relatively large region of the background, and it is therefore probable that some parts of them be similar to the underlying background thus not being detected in the change detection process. The function χ is therefore intended as a penalizing factor for blobs not correctly segmented.

Since υ and χ range from 0 to 1, the same is true for ϕ which in the end gives an estimate of the level of performance of each sensor for each extracted blob. As shown in Fig. 3, the ϕ values (reported below the bounding boxes) of the blobs extracted from the infrared

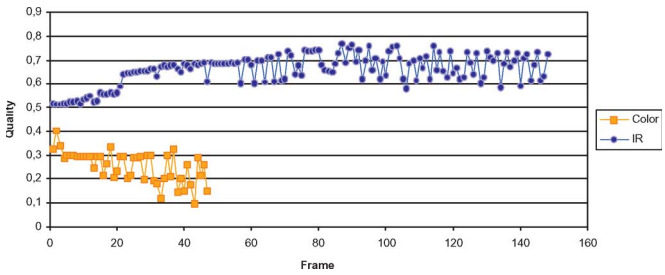

Fig. 2. ϕ values for some frames of the sequence in the presence of fog. The IR sensor clearly outperforms the color camera in detecting the target.

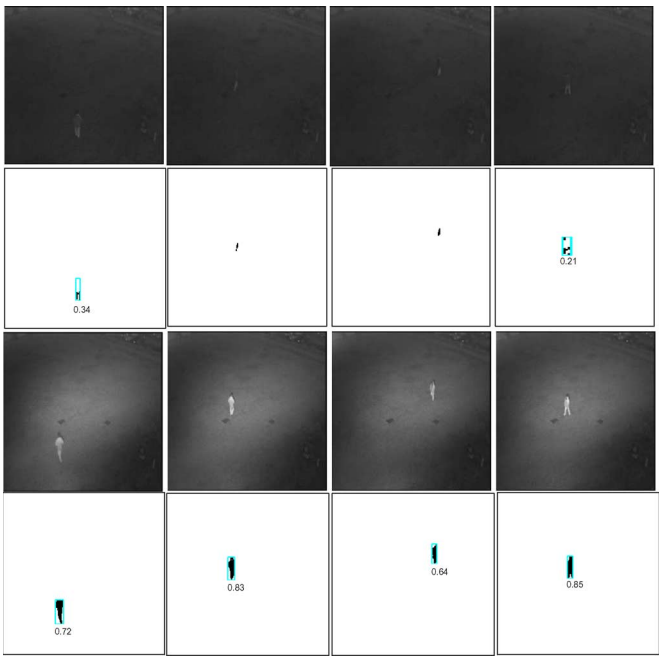

Fig. 3. Four image frames and blobs from a video sequence. Frames acquired by the color camera (top row), blobs obtained from the color camera sequence (second row), frames acquired by the IR camera (third row), blobs obtained from the IR camera sequence (fourth row).

(IR) sensor are considerably higher than those extracted from the optical one. Fig. 2 shows the ϕ values for the blob of the walking person in Fig. 3. The scene is irradiated by IR rays and monitored by a color [set to black-and-white (b/w) mode] and a b/w camera with near IR response. It can be seen how the IR sensor outperforms the color camera: the ϕ values of the blob corresponding to the IR sensor are consistently higher. This is directly reflected by the correct segmentation of the silhouette of the person. It can also be noted how the color camera is not able to discriminate the person as he moves away and into a fog bank (the ϕ values in Fig. 2 are not indicated in the graph since the blob is not detected).

The ϕ measure has to be computed for each blob detected by each sensor. This allows to compare and rate the performance of the available sensor as can be seen in the experiment reported in Fig. 4. Two color cameras have been employed to follow the movements of three persons walking in a courtyard.

The first row of Fig. 4 shows images taken from the first camera which is superior in quality to the second camera and proved to be more effective in detecting the walking persons. In fact, the first camera yields images with better contrast that benefit the change detection process. Even though the first sensor is monitoring the area with a configuration of the optics more wide-angled than the second one (thus detecting smaller blobs), it still performed slightly better. This is reflected by the ϕ values of the blobs in the second row which are generally greater than those in the fourth row. In this experiment both sensors performed reasonably well.

From this experiment, other benefits of employing multiple sensors can be highlighted in the following.

1) The system exploits the estimates of just one sensor when the other one is not giving readings (i.e., the target is out of the field of view, e.g., in Fig. 4, first column, the person on the left in rows 1–2 is not present in the field of view of the second sensor, rows 3–4).
2) Multiple views of the same target can help disambiguate situations of partial or total occlusions (second column of Fig. 4), therefore helping in maintaining a correct and continuous tracking of the targets. Note that the ϕ value is not computed for the blob detected by the first sensor since it is recognized as a compound object generated by an occlusion and therefore will not be associated to any of the three objects present at the previous time instant. Since occlusions severely spoil the appearance of a blob, its localization is generally extremely imprecise. Therefore, the ϕ value of occluded blobs is set to 0 by default.
3) Data fusion compensates for errors due to the homographic transformation from image pixels to map points. In fact, the first sensor gives better segmentation results. But, due to the wide-angle setup of the optics, homographic errors are more probable.

B. Sensor Weighting Function

The idea is to obtain from the Kalman filter a fused estimate more influenced by accurate local estimates and almost unaffected by inaccurate ones. Unreliable sensors, namely those whose detected blobs have ϕ values below a given threshold, may be even completely discarded in the fusion process. In this way, the fused estimates are obtained only from the pool of sensors that are giving an acceptable performance.

The filter's responsiveness can be adjusted through the measurement error covariance matrix $\mathbf{R}$ that at time k is given by

$$\mathbf{R}_k = \begin{pmatrix} r_{xx}^k & 0 \\ 0 & r_{yy}^k \end{pmatrix}.$$

Position measurement error is not assumed to be cross-correlated, therefore r_{xy} and r_{yx} are set to zero. If the eigenvalues of a given matrix $\mathbf{R}'$ associated with a sensor s' are smaller than those of $\mathbf{R}''$ associated with s'', the corresponding measurement will have a larger weight. ϕ values are then used to regulate the measurement error covariance matrix that eventually weights state estimates in the fusion process. Since the matrix R influences the Kalman filter gain matrix $\mathbf{K}$, it will affect the fused state estimate $\hat{x}_{k|k}$ and its corresponding error covariance matrix $P_{k|k}$, as shown in (6)–(8) in the Appendix. The following function for the position measurement error variance has been developed:

$$r_{xx}^k = r_{yy}^k = r\left(\mathbf{b}_{j,k}^s\right) = \max_r^2 \left[1 - \phi\left(\mathbf{b}_{j,k}^s\right)\right] \tag{5}$$

where $\max_r^2$ is a constant corresponding to the maximum measurement error variance (which was experimentally evaluated as 5 m^2). The function is therefore used to adjust the measurement position error so that the map positions calculated for blobs with high ϕ values are trusted more (the values of $\mathbf{R}_k$ are close to zero), while blobs poorly

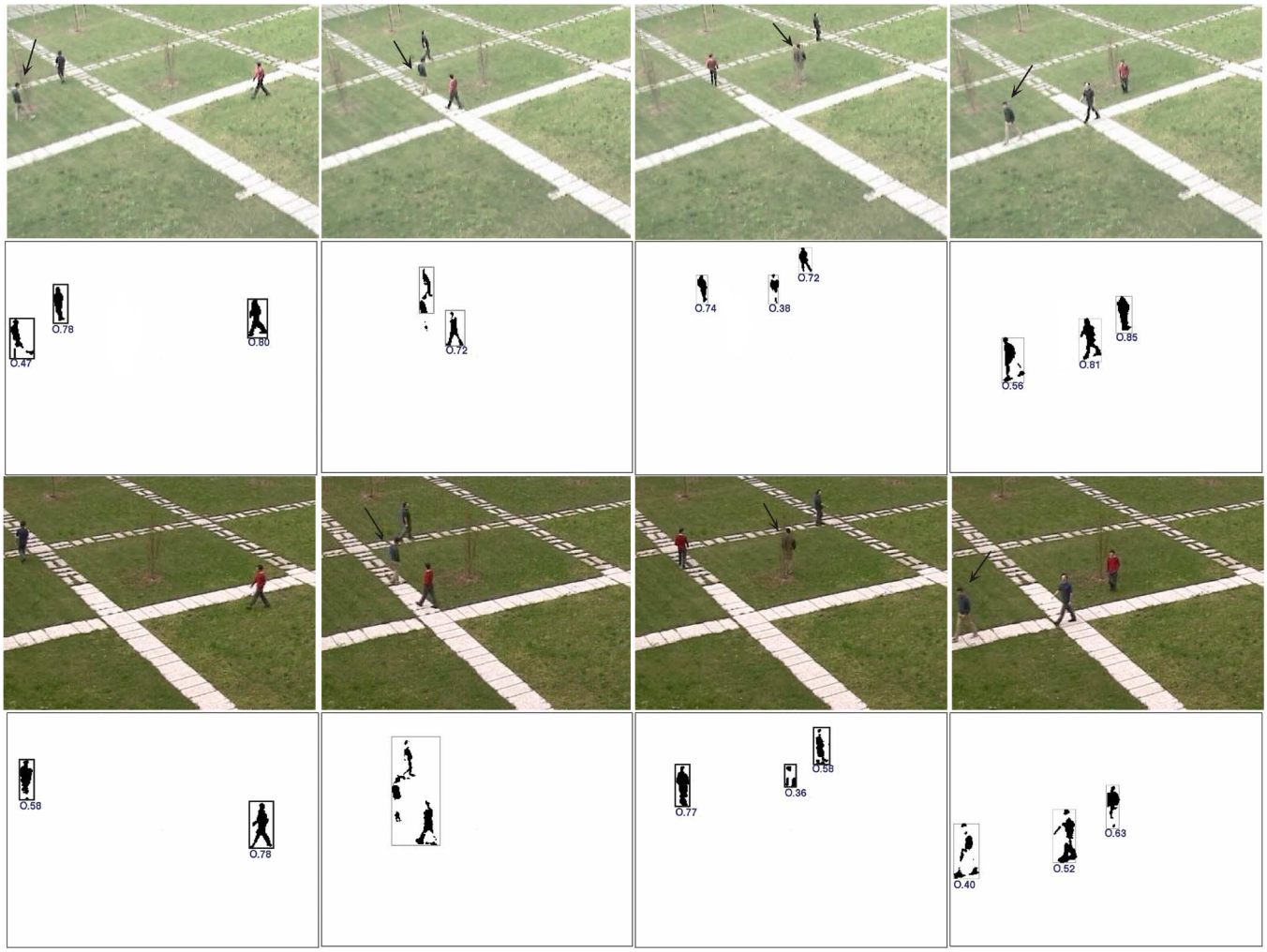

Fig. 4. Two color sensors monitoring a courtyard. The first sensor (first and second row) is performing better due to better contrast between the objects and the background. This is reflected by higher ϕ values (indicated below each blob in the binary images). Note that ϕ values are not computed when an occlusion is detected (second column).

detected (low ϕ value) are trusted less (the values of $\mathbf{R}_k$ are close to the maximum).

III. Experimental Results

Experiments with real video sequences have been carried out in order to test the performance of the proposed approach. Up to three cameras, directly attached to a processing unit, have been employed to track the movements of a person walking in outdoor scenes. Each experiment is comprised of sequences 1-h long each; the detection rate was 90% and the false alarm rate was 0.01%. A single target is certainly not critical for a vision tracking system, however the purpose here is to evaluate the accuracy of the trajectories, not the tracking algorithms. Root mean square (rms) errors of the trajectories resulting from the individual sensors, blind fusion (sensor performance is not considered), and the proposed performance-based (PB) fusion have been computed against ground truth trajectories. These were obtained by positioning markers on the ground and by timing the target. Cubic splines were then computed to interpolate the markers on the ground as a function of time.

A. Two Cameras

In this experiment, a person is walking in a parking lot and making a curved trajectory. In this case, illumination conditions are more challenging since natural light is low. As shown in Fig. 5, the second sensor is more informative and provides a brighter image. This is also confirmed by the quality assigned to the detected target: the ϕ values assigned to the target are clearly different, as shown in Fig. 6.

Fig. 5. Two-camera experiment. The second camera (b) has better sensibility and provides a brighter image.

The trajectory according to the individual sensors and the two fusion approaches are shown in Figs. 7 and 8.

It can be seen how the PB fusion approach dynamically regulates the fusion process taking into account the quality of the detected target. In fact, the fused trajectory is mostly determined by the second camera that is correctly capturing the motion of the target. The errors of the

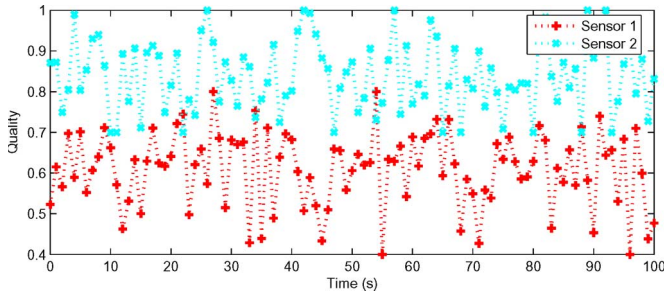

Fig. 6. Two-camera experiment. Target's quality according to sensors. The second sensor is detecting the target with higher ϕ values.

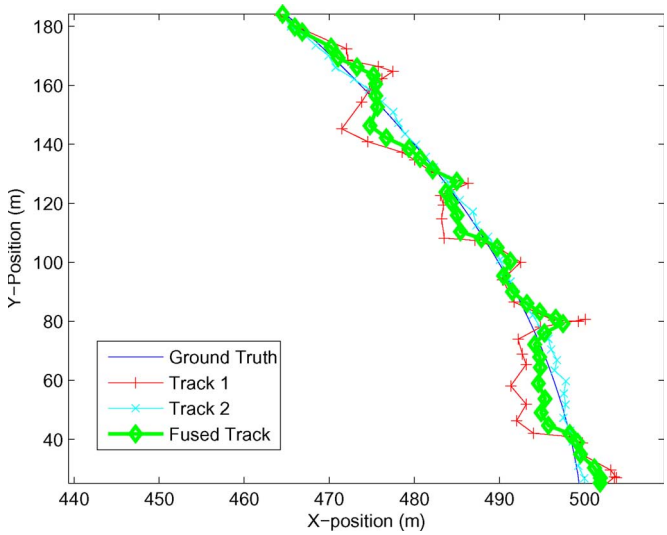

Fig. 7. Two-camera experiment. Target's trajectory according to ground truth (solid line), the two sensors, and blind fusion.

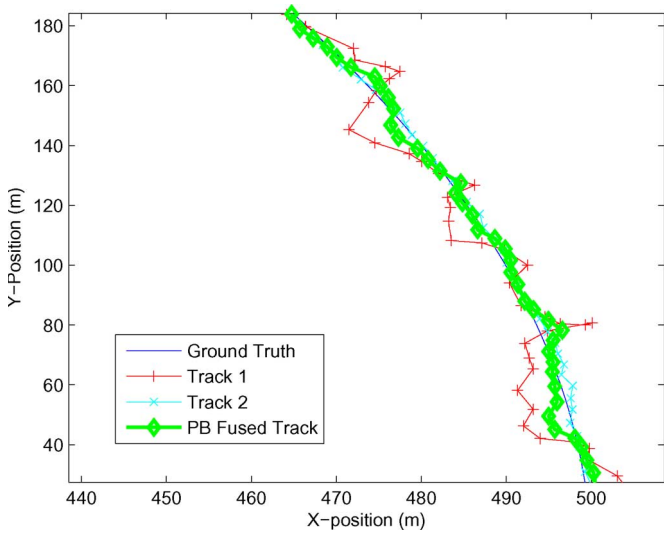

Fig. 8. Two-camera experiment. Target's trajectory obtained through PB fusion (compare with Fig. 7).

estimated tracks according to the two cameras are reported in the charts of Fig. 9, while those committed by standard and PB fusion are shown in Fig. 10.

Finally, the standard deviation of the error is reported in Table I. As shown, PB fusion is performing better than blind fusion.

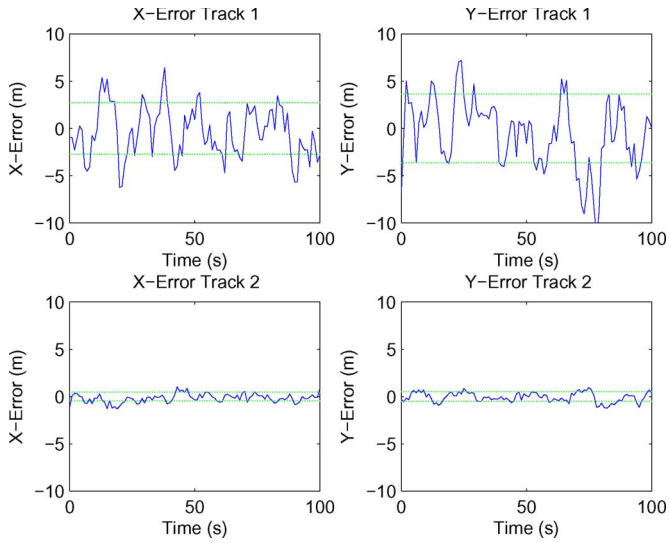

Fig. 9. Two-camera experiment. Position errors of the trajectory produced by (top) the first and (bottom) the second sensor.

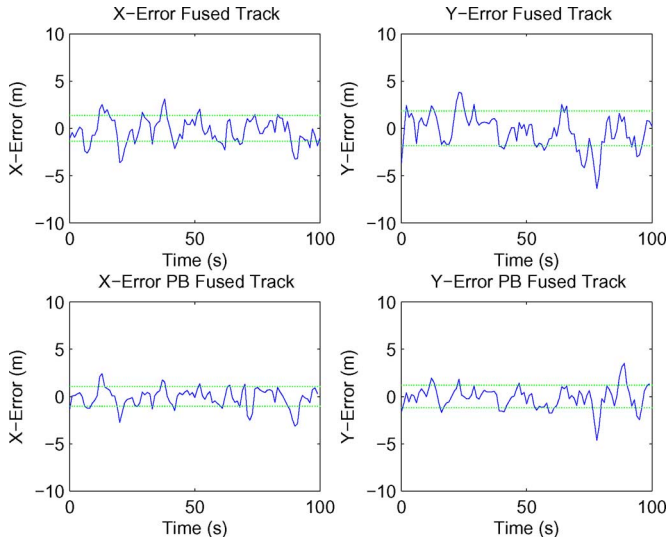

Fig. 10. Two-camera experiment. Position error of the trajectory obtained from (top) blind fusion and (bottom) the proposed approach.

TABLE I
RMS ERRORS IN THE TWO-CAMERA EXPERIMENT

	Error σ_x (m)	Error σ_y (m)
Measurement 1	3.317	4.134
Measurement 2	1.040	0.99
Track 1	2.723	3.636
Track 2	0.463	0.515
Blind Fusion	1.367	1.827
PB Fusion	1.029	1.190

B. Three Cameras

In this experiment, three cameras are observing the pathway by a parking lot in daylight (Fig. 11). Sensor 1 is a fixed b/w camera with very wide angle lens, sensor 2 is a zoomed active black-and-white camera, while sensor 3 is a fixed color camera with a wide-angle lens.

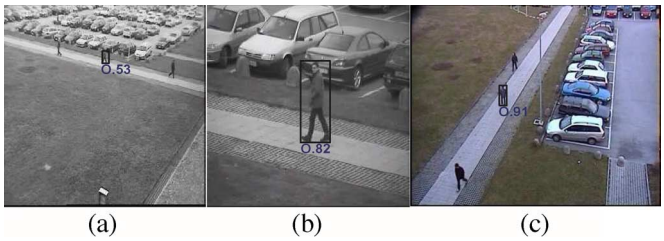

Fig. 11. Frames from the three-camera experiment. (a) First sensor is the one performing worst, (b) sensor 2 is the second best, and (c) sensor 3 is the best one.

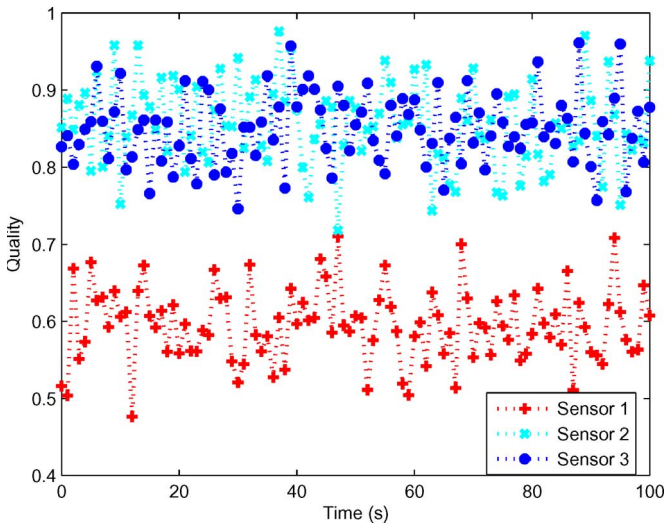

Fig. 12. Three-camera experiment. Target's quality according to three sensors. Sensors 2 and 3 are yielding higher ϕ values.

The active camera was registered on the ground plane as well [1] and was following a target selected by the operator among those detected by the two fixed cameras (three in the shown scene). For the sake of clarity, only the tracks and quality information regarding one of them will be here reported.

The sensors were performing differently, as shown in Fig. 12. In particular, while sensors 2 and 3 were giving high ϕ values, sensor 1 was performing poorly. Targets detected by sensor 1 are very small and easily corruptible by noise. This in turn translates into localization errors on the map. The quality function correctly evaluated the discriminative capability of the sensor, giving it consistently low values.

This is reflected by the detected trajectories, as shown in Fig. 13, where track 1 is considerably noisy, while tracks 2 and 3 are closer to ground truth.

Fig. 13 also shows the trajectories obtained by fusing two and three sensors. In the former case, sensors 1 and 2 were fused. It can be clearly seen how the fused trajectory benefits from the addition of track 3, i.e., the fused trajectory of the three sensors is closer to the ground truth than the fusion of the first two alone. Fig. 14 shows how the proposed PB fusion process brings the two fused trajectories even closer to the ground truth. This is numerically confirmed by the error plots in Figs. 15–17.

Table II summarizes the results of this experiment. It can be seen how the blind fusion of the three sensors is better than the blind fusion of sensors 1 and 2, and how PB fusion is better than blind fusion both in the two- and three-sensors case. Note that, in this case, the fusion of the three sensors outperforms the individual sensors.

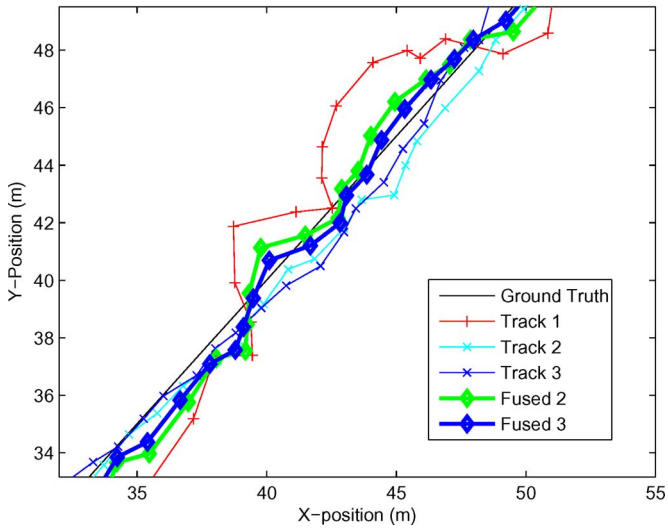

Fig. 13. Target's trajectory according to ground truth (solid line), the three sensors, and blind fusion.

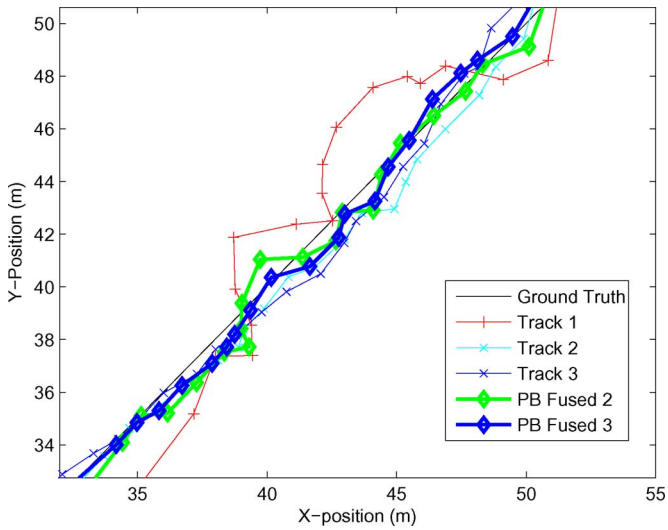

Fig. 14. Single tracks and PB fusion.

C. Discussion

In summary, the advantages of using adaptive fusion for the estimation of trajectories in a surveillance application are the following.

1) The system automatically takes into consideration redundant data when available.
2) The presence of two points of view can help disambiguate situations of partial or total occlusions, therefore maintaining a correct and continuous track of the targets.
3) There is an explicit weighting of the measurements in the fusion process through the ϕ function to account for segmentation errors, and hence localization errors caused by segmentation errors.
4) Data fusion increases the confidence in estimates.

IV. Conclusion

In this correspondence, the problem of improving tracking accuracy through multiple visual sensors in a distributed framework has been

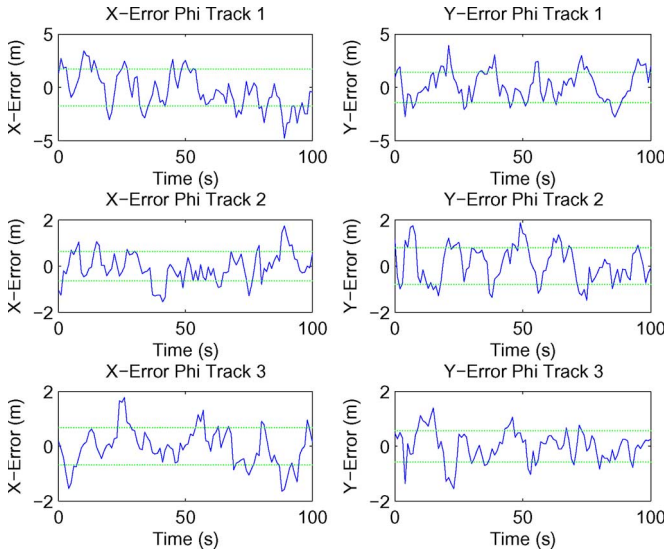

Fig. 15. Three-camera experiment. Position errors of (top) the first, (middle) second, and (bottom) third sensor.

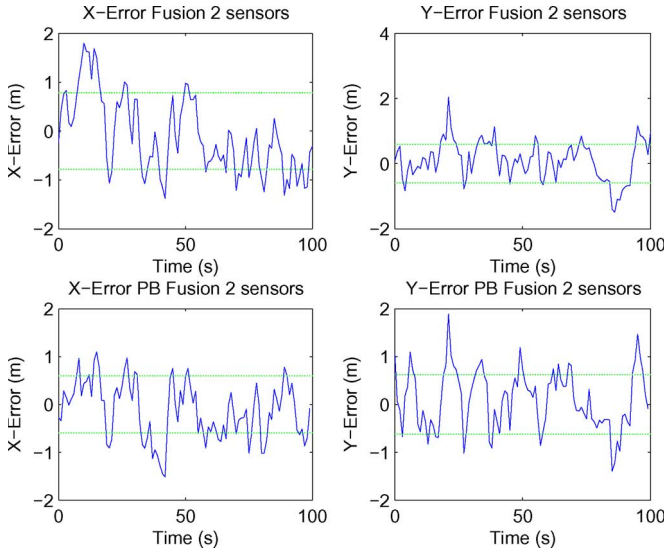

Fig. 16. Three-camera experiment. Position error of (top) the first, (middle) the second, and (bottom) the third sensor.

addressed for an outdoor video surveillance application. A data fusion approach has been proposed to adaptively combine, according to the performance of each sensor, the position of a target resulting in a unified estimate. Sensor reliability is explicitly considered and a confidence function has been defined to automatically weight redundant estimates of the location of the targets in the fusion process. In this way, localization errors due to incorrect segmentation of the blobs have been reduced, as well as the errors due to homographic transformations. Experimental results have shown the effectiveness of the proposed approach in terms of tracking accuracy in comparison with single-camera systems.

Future development of the presented approach is expected to proceed in at least two directions. The first one is to introduce additional factors to the quality function ϕ described here. It could be extended to include other elements such as distance of the blob from the camera, or global illumination of the scene. The second possible direction is to experiment with other methods and filtering techniques for tracking such as particle filters.

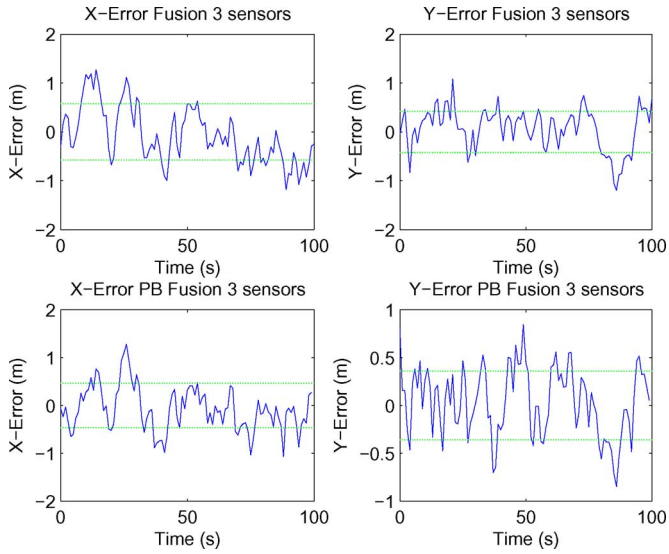

Fig. 17. Three-camera experiment. Position error of (top) of blind and (bottom) of PB fusion of the three sensors.

TABLE II
RMS ERRORS IN THE THREE-CAMERA EXPERIMENT

	Error σ_x (m)	Error σ_y (m)
Measurement 1	2.273	2.022
Measurement 2	1.137	1.178
Measurement 3	0.997	1.030
Track 1	1.541	1.238
Track 2	0.528	0.525
Track 3	0.576	0.512
Blind Fusion 2 sensors	0.782	0.623
PB Fusion 2 sensors	0.595	0.591
Blind Fusion 3 sensors	0.575	0.4220
PB Fusion 3 sensors	0.463	0.358

APPENDIX
FUSION PROCESS

The fused state $\hat{\mathbf{x}}_{k|k}$ for a given target is obtained fusing all the local estimates matching the predicted fused state $\hat{\mathbf{x}}_{k|k-1}$. In the case of two estimates $\hat{\mathbf{x}}^i_{k|k}$ and $\hat{\mathbf{x}}^j_{k|k}$ from sensors i and j, fusion is performed as follows:

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}^i_{k|k} + \left[\mathbf{P}^i_{k|k} - \mathbf{P}^{ij}_{k|k}\right]\left[\mathbf{P}^i_{k|k} + \mathbf{P}^j_{k|k} - \mathbf{P}^{ij}_{k|k} - \mathbf{P}^{ji}_{k|k}\right]^{-1} \times \left(\hat{\mathbf{x}}^j_{k|k} - \hat{\mathbf{x}}^i_{k|k}\right) \quad (6)$$

where $\mathbf{P}^i_{k|k}$ and $\mathbf{P}^j_{k|k}$ are the error covariance matrices for the local estimates and $\mathbf{P}^{ij}_{k|k} = (\mathbf{P}^{ji}_{k|k})^{\mathrm{T}}$ is the cross-covariance matrix, which is given by the following recursive equation:

$$\mathbf{P}^{ij}_{k|k} = \left[\mathbf{I} - \mathbf{K}^i_k\mathbf{H}^i_k\right]\left[\mathbf{F}_{k-1}\mathbf{P}^{ij}_{k-1|k-1}\mathbf{F}^{\mathrm{T}}_{k-1} + \mathbf{\Gamma}_{k-1}\mathbf{Q}\mathbf{\Gamma}^{\mathrm{T}}_{k-1}\right] \times \left[\mathbf{I} - \mathbf{K}^j_k\mathbf{H}^j_k\right] \quad (7)$$

where $\mathbf{K}_k^s$ is the Kalman filter gain matrix for sensor s at time kT, $\mathbf{\Gamma}_k$ is the process noise matrix at time kT, and $\mathbf{Q}$ is the process noise covariance matrix. Once the cross-covariance matrix is available, the covariance matrix associated with the fused estimate can be evaluated as follows:

$$\mathbf{P}_{k|k} = \mathbf{P}^i_{k|k} - \left[\mathbf{P}^i_{k|k} - \mathbf{P}^{ij}_{k|k}\right] \times \left[\mathbf{P}^i_{k|k} + \mathbf{P}^j_{k|k} - \mathbf{P}^{ij}_{k|k} - \mathbf{P}^{ji}_{k|k}\right]^{-1} \left[\mathbf{P}^i_{k|k} - \mathbf{P}^{ji}_{k|k}\right]. \quad (8)$$

Further details can be found in [13] and [14], and generalized fusion equations for the case of N sensors can be found in [15].

REFERENCES

[1] G. L. Foresti, C. Micheloni, L. Snidaro, P. Remagnino, and T. Ellis, "Active video-based surveillance systems," *IEEE Signal Process. Mag.*, vol. 22, no. 2, pp. 25–37, Mar. 2005.

[2] G. L. Foresti, C. S. Regazzoni, and P. K. Varshney, *Multisensor Surveillance Systems: The Fusion Perspective*. Norwell, MA: Kluwer, 2003.

[3] D. Makris and T. Ellis, "Learning semantic scene models from observing activity in visual surveillance," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 3, pp. 397–408, Jun. 2005.

[4] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation With Applications to Tracking and Navigation*. New York: Wiley-Interscience, Jun. 2001.

[5] D. L. Hall and S. A. McMullen, *Mathematical Techniques in Multisensor Data Fusion*. Norwood, MA: Artech House, Mar. 2004.

[6] J. Triesch and C. von der Malsburg, "Democratic integration: Self-organized integration of adaptive cues," *Neural Comput.*, vol. 13, no. 9, pp. 2049–2074, 2001.

[7] O. Kahler, J. Denzler, and J. Triesch, "Hierarchical sensor data fusion by probabilistic cue integration for robust 3D object tracking," in *Proc. 6th IEEE Southwest Symp. Image Anal. and Interpret.*, Lake Tahoe, NV, Mar. 2004, pp. 216–220.

[8] L. Snidaro, G. L. Foresti, R. Niu, and P. K. Varshney, "Sensor fusion for video surveillance," in *Proc. 7th Int. Conf. Inf. Fusion*. Stockholm, Sweden: Int. Soc. Inf. Fusion, Jun. 2004, vol. 2, pp. 739–746.

[9] I. Avcibaş, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," *J. Electron. Imaging*, vol. 11, no. 2, pp. 206–223, 2002.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[11] P. L. Correia and F. Pereira, "Objective evaluation of video segmentation quality," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 186–200, Feb. 2003.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York: Cambridge Univ. Press, Mar. 2004.

[13] Y. Bar-Shalom and X. R. Li, *Multitarget-Multisensor Tracking: Principles and Techniques*. Storrs, CT: YBS Publishing, 1995.

[14] J. B. Gao and C. J. Harris, "Some remarks on Kalman filters for the multisensor fusion," *Inf. Fusion*, vol. 3, no. 3, pp. 191–201, Sep. 2002.

[15] H. Chen, T. Kirubarajan, and Y. Bar-Shalom, "Performance limits of track-to-track fusion versus centralized estimation: Theory and application," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 2, pp. 386–400, Apr. 2003.