Fusion of Non-Visual Modalities Into the Probabilistic Occupancy Map Framework for Person Localization

Rok Mandeljc, Janez Perš, Matej Kristan and Stanislav Kovačič
Faculty of Electrical Engineering
University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
Email: rok.mandeljc@fe.uni-lj.si

***Abstract*—In this paper we investigate the possibilities for fusion of non-visual sensor modalities into state-of-the-art vision-based framework for person detection and localization, the Probabilistic Occupancy Map (POM), with the aim of improving the frame-by-frame localization results in a realistic (cluttered) indoor environment. We point out the aspects that need to be considered when fusing non-visual sensor information into POM and provide a mathematical model for it. We demonstrate the proposed fusion method on the example of multi-camera and radio-based person localization setup. The performance of both systems is evaluated, showing their strengths and weaknesses. We show that localization results may be significantly improved by fusing the information from the radio-based system into the camera-based POM framework using the proposed model.**

I. Introduction

In recent years, the problem of object detection, localization and tracking has received lots of attention. This interest coincides with the demand for such information, coming from a wide spectrum of applications that have emerged in diverse fields such as manufacturing, military, surveillance and security, transport and logistics, medical care, childcare and tracking in sports. Various localization solutions based on different sensor modalities have been proposed [1], [2]. Two strong research areas are tracking by video cameras and localization based on radio technologies. Lately, sensor fusion has gained prominence as a paradigm for overcoming limitations of the individual sensor modalities [1].

One of the popular methods for sensor information fusion is an occupancy map. Often employed in the robotics community [3], the occupancy map (occupancy grid) involves dividing the area of interest into a grid and estimating the probabilities of the cells' occupancy, given the sensor data. This is done under the independence assumption, which allows the problem of estimating the posterior probability of the whole map to be broken down into individual binary problems of estimating the occupancy probability for each cell. Occupancy maps allow efficient data aggregation, with each sensor contributing to the occupancy likelihood of every cell.

In this paper, we investigate the possibilities for fusion of non-visual sensor modalities into the Probabilistic Occupancy Map [4] framework (POM), which we consider a state-of-the-art vision-based framework for person detection and localization. The aim of our investigation is to improve, by means of fusion, the frame-by-frame localization performance, for example in a realistic (cluttered) indoor environment. In such an environment the people are occluded not only by other people (the situation POM handles by design), but also by objects such as office furniture (e.g. desk and chairs). This causes difficulties in accurate and reliable position estimation. These problems can be alleviated with fusion of an additional, complementary sensor modality, which may have difficulties of its own in a cluttered environment.

To the best of our knowledge, so far no attempt has been done to extend the POM algorithm to fuse additional (non-visual) sensory modalities within its framework.

Our main contribution is therefore the extension of the original POM formulation to include arbitrary sensor modalities (Section III). By showing how the original model for visual sensors fits into it, we gain insight about the aspects that need to be taken into consideration when integrating arbitrary sensor modalities into POM. The second contribution is our application, described in Section IV, where we evaluate the performance of two localization systems in a cluttered environment; the first system consists of four calibrated cameras and uses the POM algorithm, while the second system is radio-based. We point out strengths and weaknesses of both systems by evaluating the 2-D position error and the number of false positive and false negative detections. Finally, taking advantage of both systems' strengths, we fuse the information from the radio-based system into POM in order to improve its frame-by-frame localization results.

II. Related work

Object detection and tracking using video cameras has a long tradition, and many different approaches have been proposed [5]. Much of the recent research has been done in the context of surveillance and tracking of players in sports [6]. In order to cope with occlusions and complex scenes, multi-camera based approaches, which fuse information from multiple views, have been proposed [7], [8], [9].

The occupancy map was introduced in the field of computer vision with works of Beymer [10] and Yang *et al.* [11], both in the context of estimating the number of people in a room. In the work of Franco and Boyer [3], an occupancy map is employed as a framework for multi-view silhouette cue fusion.

978-1-4577-1707-9/11/$26.00 ©2011 IEEE

These are all bottom-up approaches; they project the points of the foreground likelihood (stereo disparities or background subtracted regions) of each view into the ground plane and then infer the occupancy likelihood from the amount of projected points in each cell. The recent works of Delannay *et al.* [12] and Muñoz *et al.* [13] also fall into this category.

The Probability Occupancy Map framework (POM) by Fleuret *et al.* [4], on the other hand, employs a top-down approach; the probability of occupancy at each cell in the grid is estimated based on the back-projection of a generative model into each of the calibrated views and the probability field is iteratively optimized from initial values until the back-projections match the input foreground mask. POM is concerned only with person detection and localization, on frame-by-frame basis and without imposing temporal continuity; it can, however, be incorporated into a tracking framework, as shown in [4]. In [14], the POM framework is used in conjunction with people detectors. Due to its iterative nature and the use of back-projection, the POM framework implicitly handles the occlusions between people, and is considered state-of-the-art.

Another top-down approach utilizing the occupancy map is the work of Alahi *et al.* [15], where the occupancy map is estimated from the background subtraction results via generative model and imposed sparsity constraint.

Very recently, Lee *et al.* [16] demonstrated, using their own framework, the benefits of combining both the results of background subtraction and the output of a human detector.

Fusion of visual information with other sensor modalities is also an active research topic. There have been various attempts at the person localization via fusion of the computer vision and radio techniques, such as RFID [17] or WiFi signal strength [17], [18], [19]. Town [20] describes a sentient office, in which the visual information from calibrated cameras is combined with the location events from an ultrasonic tracking system.

A very prominent area of multi-modal sensor fusion is fusion of audio-visual information [21]; for this task, Bayesian networks with particle-filtering-based inference are commonly used. These approaches involve temporal filtering, and therefore their results are not directly comparable to the results of the POM algorithm, which is concerned only with frame-by-frame detection and localization.

III. Probabilistic occupancy map

Probabilistic occupancy map (POM) [4] is a framework for multi-camera people detection and localization. It iteratively estimates the marginal probabilities of individuals' presence at every cell in the occupancy map of the area of interest on frame-by-frame basis.

The algorithm employs a simple appearance model, under which the silhouettes of the individuals in each of the calibrated views are approximated by a family of rectangles. These are used in each iteration step to generate, based on the currently estimated probability field, per-view synthetic images, which are compared to the input binary images that were obtained by background subtraction. The probability field is iteratively optimized, so that the generated synthetic images match the input binary images, until convergence to a fixed point is reached.

POM operates on frame-by-frame basis, without imposing temporal continuity and is, due to the back-projections and iterative nature, able to effectively handle occlusions at each frame independently.

In order to investigate the fusion of additional non-visual sensors into POM, we generalize the original formulation and show how the original model for visual cameras fits into it. Doing so we gain insight about the properties of the algorithm that need to be taken into consideration when integrating additional sensor modalities.

A. General formulation

Given G cells on the ground plane, we denote by $\mathbf{X}$ a vector of binary random variables $(X_1, \ldots, X_G)$ standing for the occupancy of each cell ($X_k = 1$ if cell k is occupied and $X_k = 0$ if it is vacant). The goal of POM is to estimate the probabilities of occupancy for each cell, $P(X_k = 1 \mid \mathbf{B})$, given the information from C sensors $\mathbf{B} = (B_1, \ldots, B_C)$. The occupancy probabilities are approximated by the marginals of a product law Q, which minimizes the Kullback-Liebler divergence from the "true" posterior $P(\mathbf{X} \mid \mathbf{B})$.

The expression for the marginal probability $q_k = Q(X_k = 1)$, is derived in [4] via minimization of Kullback-Liebler divergence. Here, we write the expression in a more general form,

$$q_k = \frac{1}{1 + e^{\lambda_k + \sum_c R_c(k,\mathbf{X})}}, \tag{1}$$

where λ_k is the log-ratio of the prior probability, and $R_c(k, \mathbf{X})$ is a term that encapsulates all the information from the sensor c about the occupancy of the cell k,

$$\begin{aligned} R_c(k,\mathbf{X}) = E_Q\{\log P(B_c \mid \mathbf{X}) \mid X_k = 0\} - \\ - E_Q\{\log P(B_c \mid \mathbf{X}) \mid X_k = 1\}, \end{aligned} \tag{2}$$

where E_Q denotes the expectation under the approximation $\mathbf{X} \sim Q$. In the original formulation, B_c stands for the binary image obtained by background subtraction in camera c; in the above formulation, sensor c can be of an arbitrary modality.

B. Visual sensors

For visual sensors (cameras), [4] presents a generative model that relates the values of X_k to the images produced by background subtraction. The authors define a normalized pseudo-distance between the background subtraction image B_c and a synthetic image A_c, which is generated at each iteration step from the currently estimated values of $\mathbf{X}$. The conditional distribution is modeled as $P(B_c \mid \mathbf{X}) = \frac{1}{Z} e^{-\Psi(B_c, Ac)}$, where Ψ denotes the pseudo-distance.

In order to make the problem solvable, the approximation $E_Q\{\Psi(B_c, A_c) \mid X_k = \xi\} \approx \Psi(B_c, E_Q(A_c \mid X_k = \xi))$, with $\xi = \{0, 1\}$, is made under the assumptions that are

TABLE I

SINGLE CAMERA $\mathcal{R}_c$ RATIO VALUES FOR AN OCCUPIED AND A VACANT CELL DURING POM'S CONVERGENCE.

Iteration	**1**	**2**	**5**	**12**	**16**
Occupied	1.3e+04	1.07	1.76	12.75	26.18
Vacant	1.47	0.84	0.75	0.51	0.23

Iteration	**20**	**30**	**40**	**50**	**62 (final)**
Occupied	84.30	1.7e+03	3.2e+03	3.5e+03	3.5e+03
Vacant	0.03	1.2e-05	8.4e-07	5.7e-07	5.5e-07

detailed in [4]. The term (2) for the background-subtraction-based sensor can therefore be written as

$$R_c(k, \mathbf{X}) = \log \frac{e^{-\Psi(B_c, E_Q\{A_c | X_k = 0\})}}{e^{-\Psi(B_c, E_Q\{A_c | X_k = 1\})}} = \\ = -\log \mathcal{R}_c(k, \mathbf{X}). \tag{3}$$

The conditional synthetic images $E_Q(A_c \mid X_k = \xi)$ correspond to the average synthetic image $E_Q\{A_c\}$ with q_k forced to 0 and 1, respectively; the first case represents the hypothesis that the cell k is vacant, and the second that the cell k is occupied. If the hypothesis that the cell k is occupied improves the fit of the synthetic image to the input image for camera c, the ratio $\mathcal{R}_c$ has a large value, and consequently the term R_c is negative. As can be seen from (1), a negative R_c leads to a larger q_k. Similarly, if the hypothesis that the cell is vacant improves the fit, the ratio $\mathcal{R}_c$ goes towards zero, R_c is positive and therefore q_k decreases.

An important aspect of this model is that the synthetic images at a given iteration step are generated from the currently estimated occupancy probabilities. Therefore they are, along with corresponding R_c and $\mathcal{R}_c$, their functions, and change with each iteration step. This, on one hand, allows POM to implicitly handle the occlusions; on the other hand, it poses a difficulty for integration of sensors with different models. Table I shows the $\mathcal{R}_c$ values for one of the cameras during POM's convergence, both for an occupied and a vacant cell.

At this point we should also note that this model is not limited only to binary images obtained from visual cameras by background subtraction. It can also be applied to other kind of information obtained from visual cameras (e.g. people detectors [22], [23]), or even to cameras with other sensing modalities (e.g. infrared cameras). The only requirement is that the information can be interpreted as binary images where the blobs correspond to the individuals' silhouettes. More importantly, since the same model is used, such information can be seamlessly fused within the POM framework.

C. Model for non-visual sensor modalities

Because non-visual sensor modalities require a different model than the visual sensors (different types of cameras), their fusion within the POM framework can prove difficult. One of the obstacles are the expectations E_Q in (2); their evaluation may be intractable, thus requiring the appropriate approximations to be made.

Even so, difficulties may arise due to POM's iterative nature. If the value range of non-visual modality's R_s[1] term is not comparable to the value range of the cameras' R_c terms (see Table I), one or the other modality consistently prevails. Furthermore, the values of the R_c terms change at each iteration step; if R_s is also modeled as a function of the currently estimated probabilities, it needs to change at a comparable rate. In practice, this is hard to achieve.

Therefore we model the conditional probabilities in the non-visual sensor's R_s term as being independent of the currently estimated probabilities and investigate the model's properties. With such a model, the expectations E_Q cancel out; term (2) becomes

$$R_s(k) = \log \frac{P(B_s \mid X_k = 0)}{P(B_s \mid X_k = 1)} = -\log \mathcal{R}_s(k), \tag{4}$$

and can be interpreted as a log-ratio of conditional probabilities of sensor making the observation at hand, given that the cell k is occupied or vacant, respectively.

Since the value of R_s is constant for given k throughout all the iterations, it influences the POM's convergence when its values are of greater or comparable order of magnitude to the values of cameras' R_c terms (see Table I). In practice, the value of the R_s term should be small (comparable to R_c values in the initial iteration steps), and influence the convergence only in the initial iteration steps. If the R_s is significantly larger than R_c terms, it causes a peak in the estimated probability for the given cell. This results in the saturation of the cameras' synthetic images, which causes the algorithm's convergence to end in its early steps.

Therefore in such a fusion scheme, the non-visual sensor has an auxiliary role; the majority of information required for the algorithm's convergence is provided by cameras, while the auxiliary sensor information increases the robustness. In cases when the ambiguity of the visual information causes the algorithm to favor the convergence to an incorrect solution, yet there is also visual evidence for the correct solution, the additional information is used to change this tendency at the initial iteration steps, causing the algorithm to converge to the correct solution. At the later iteration steps, as the magnitude of the R_c terms increases, the constant R_s term becomes irrelevant; however at this point the algorithm is already converging towards the correct solution.

On the other hand, even if auxiliary sensor's information is at times unreliable, it does not disrupt the algorithm's convergence towards the correct solution, as long as the visual information prevails in the early iteration steps. If both cameras and the auxiliary sensor fail to provide sufficient information for convergence to the correct solution, the algorithm will fail; however, such cases can be reliably handled only by imposing temporal continuity.

IV. APPLICATION EXAMPLE

In this section, we demonstrate the proposed fusion of a non-visual sensor modality into POM on the example of

[1] In the remainder of the paper we use subscript s to denote the non-visual sensor and subscript c to denote one of the visual sensors. Therefore, R_s is an equivalent of the R_c term (2), but for the non-visual sensor.

person localization in a cluttered environment. We employ two of the most prominent localization technologies — localization using visual cameras (and POM) and a radio based localization system. We evaluate their strengths and weaknesses and demonstrate the benefits of their fusion.

Localization in a cluttered environment poses a challenge both for camera-based and radio-based approaches. On one hand, objects such as the office furniture occlude portions of subjects' bodies, causing difficulties in accurate and reliable camera-based position estimation. On the other hand, the radio-based localization is error-prone due to the presence of radio-reflective (metallic) objects, which cause multipath-related problems in the radio-based range estimation.

A. Ubisense Ultra-Wideband radio localization system

For radio-based localization, we use the Ubisense Real-Time Localization System [24], which is based on the Ultra-Wideband (UWB) radio technology [25].

The system comprises a network of time-synchronized sensors (receivers) and tags (emitters) placed on the objects to be tracked. The tag's position is determined using the Time Difference of Arrival (TDoA) and the Angle of Arrival (AoA) measurements, which are combined using a least-squares algorithm [2]. The advertised accuracy of the system is 15 *cm*, with 99% of errors being within 30 *cm*; however, our preliminary experiments indicate a lower performance in a cluttered environment due to obstacles and signal reflections. The system provides the tags' position information, along with the estimations of a tag position's standard error.

As we do not have access to the raw (TDoA and AoA) measurements from the Ubisense sensors, we treat the whole system as a single sensor. We use the following model for fusion of the Ubisense position information into POM:

$$\mathcal{R}_s(k) = \alpha \cdot \max_t \{G_t(x_k, y_k)\} + \beta, \tag{5}$$

where G_t is a Gaussian centered on the coordinates of tag t detection (x_t, y_t) and evaluated at the center of the cell k, (x_k, y_k). σ_t is the standard error of a tag's position as estimated by the Ubisense system:

$$G_t(x_k, y_k) = \frac{1}{2\pi\sigma_t^2} e^{-\frac{(x_k - x_t)^2 + (y_k - y_t)^2}{2\sigma_t^2}} \tag{6}$$

Factors α and β are used to scale the Gaussian so that $\mathcal{R}_s$ values are of the same order of magnitude as the values of cameras' $\mathcal{R}_c$ ratios in the initial steps of POM's convergence. The values of both factors were determined experimentally ($\alpha = 12$, $\beta = 0.5$).

B. Experimental setup

The experiment was performed in a 7.1×6.9 *m* room, with four Axis 207W IP cameras and four Ubisense sensors placed in the corners at the height of about 2 *m*. The cameras' coverage of the room is shown in Figure 1; as can be seen, most of the locations are covered by two cameras.

Three lines were marked on the floor for the individuals to walk on (red lines in Figure 1) and a six-minute sequence involving three individuals simultaneously walking around the room was captured.

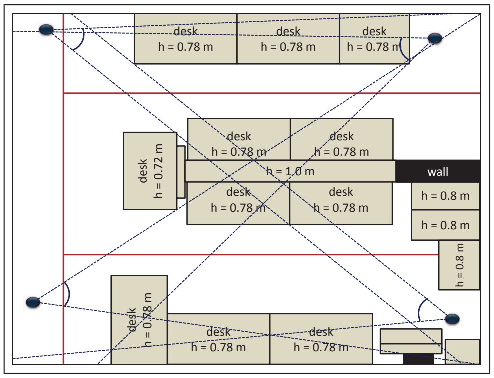


Fig. 1. The experimental setup and cameras' coverage of room (after lens distortion correction). Red lines mark the paths that the three individuals walked on during the experiment.

C. Visual cameras and background subtraction

The video from cameras (640×480 at 25 FPS) was first undistorted using the lens distortion correction [26]. The binary images were obtained by adaptive background subtraction algorithm [27], implemented in the OpenCV library. The default parameters were used.

D. POM algorithm configuration

For the POM, the room was discretized into a 20×20 *cm* grid and the rectangles corresponding to the person height of 1.8 *m* were computed using the camera calibration data. A typical person is wider than 20 *cm* and therefore fits into multiple cells; however, multiple adjacent detections could be merged during the post processing.

The reference POM algorithm [28] was modified to integrate the Ubisense position data using the proposed method. All the parameters for the POM were left at the default values, with the exception of the maximum number of iterations having been increased to 200.

During the evaluation, we consider a cell to be occupied when the posterior probability of presence as estimated by the POM is greater than 0.5.

E. Performance metric

The ground truth data was obtained by manual annotation of every 25th frame. Since the movement of individuals is relatively slow, the ground truth positions in the intermediate frames were obtained by interpolation. This way, we obtained 17503 ground truth positions in 8751 frames.

For the evaluation of a system's localization performance we calculate the distances between the estimated and the ground truth 2-D positions. From the error cumulative density function (CDF), we read the error boundaries that cover 95% and 99% of all errors and use this value as a system's performance metric. In addition, we also consider the number of false positive and false negative detections, which were determined as follows.

With the background subtraction-based POM algorithm, the identity of a person occupying a cell is unknown. Therefore, we assume that each detection belongs to the closest ground truth point and vice versa. A distance of 1 *m* was chosen to determine whether a detection is an inlier or not. If the distance from a given detection to the closest ground truth point is greater than the threshold, the detection has no corresponding ground truth point; it is *a phantom* (a false positive).

Similarly, if the distance from a ground truth point to the closest detection is greater than the threshold, the ground point has no corresponding detection and is considered *a missing detection* (a false negative). When both a phantom and a missing detection occur in a frame, we assume they correspond to each other and consider them separately as *an inaccurate detection*.

In the case of Ubisense, the identity is explicitly given; a tag can be detected only if it is present in a room, and therefore no phantom detections can occur. A missing detection occurs when the signal from a tag cannot reach the receivers and an inaccurate detection occurs when the signal arrives to the receivers via reflections due to the direct signal path being blocked; for a consistent comparison with the results of POM, we consider a detection to be inaccurate if its distance to the corresponding ground truth point is greater than 1 *m*.

F. Results and discussion

As can be seen from the summarized results in Table II and the plots in Figure 2a, the error CDF curve for camera-based POM reaches the 95% error boundary sooner than the Ubisense's CDF, whereas at the 99% error boundary, the situation is reversed. This can be explained by the gross errors in the POM detections — the phantoms and the inaccurate detections. If those outliers are discarded, the camera-based POM's error CDF reaches both error boundaries sooner than the Ubisense's (Figure 2b).

Therefore, we conclude that while the Ubisense system has a lower position accuracy than the localization with camera-based POM, it is more reliable in terms of the gross errors. Conversely, camera-based POM offers higher accuracy, but at the price of the reduced reliability.

By fusing the Ubisense position data into the POM, a significant portion of the phantom and inaccurate detections are resolved, which results in a much steeper CDF curve in Figure 2a. The error CDF curve for the inliers only (Figure 2b) shows only a marginal improvement. On one hand, this is an indication that the prevalent benefit of the fusion is indeed the resolution of the POM's gross errors. On the other hand, it also indicates that the proposed fusion scheme does not impair the correct detections, even in cases when Ubisense's position estimation is inaccurate.

Table III summarizes the number of the phantom, missing and inaccurate detections. Again, it can be seen that a significant portion of the phantom and inaccurate detections from camera-based POM are resolved by the proposed fusion method (63% and 85%, respectively). The number of the missing detections is also slightly lowered (by 32%). However, the missing detections mostly occur in cases of absence of the visual information in one of the cameras (occlusions or poor background subtraction results) and cannot be corrected even with the additional information from the Ubisense. Nor should it, because otherwise the fusion would introduce new phantoms when Ubisense's position estimation is inaccurate.

TABLE II
ERROR STATISTICS FOR THE UBISENSE RADIO TRACKING SYSTEM, CAMERA-BASED POM AND THEIR FUSION.

	Ubisense radio tracking	Camera-based POM	Fused
mean [*m*]	0.32	0.22	0.19
std [*m*]	0.34	0.25	0.18
error boundaries			
95% [*m*]	0.73	0.48	0.40
99% [*m*]	1.07	1.71	0.61
error boundaries (inliers only)			
95% [*m*]	0.68	0.42	0.39
99% [*m*]	0.87	0.64	0.54

TABLE III
PHANTOMS, MISSING AND INACCURATE DETECTIONS FOR THE UBISENSE RADIO TRACKING SYSTEM, CAMERA-BASED POM AND THEIR FUSION.

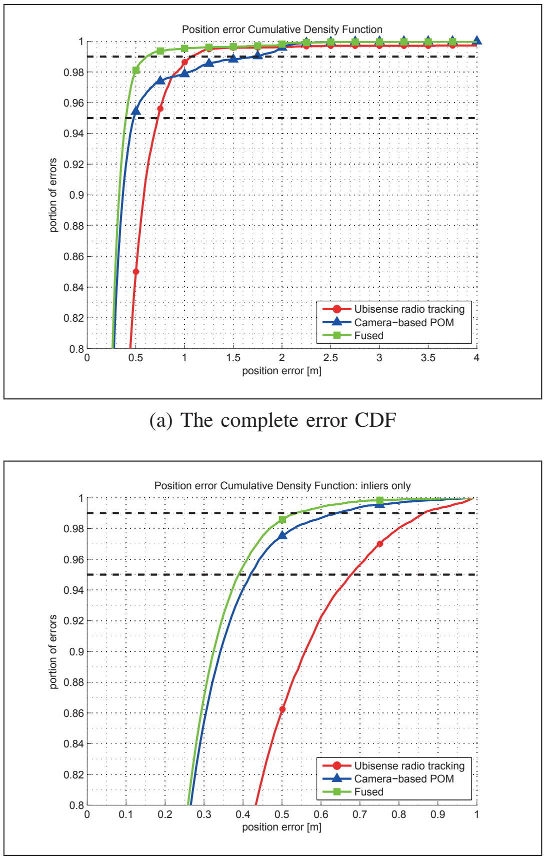
	Ubisense radio tracking	Camera-based POM	Fused
phantom	/	248	91
missing	171	1636	1114
inaccurate	248	340	49

V. CONCLUSION

We presented an approach for fusion of non-visual sensor modalities into the POM framework, with the aim of improving the localization performance in a realistic indoor environment. We extended the original formulation of the POM algorithm to include arbitrary sensor modalities and proposed a mathematical model for fusion of their information. As the original model for background-subtraction-based input images depends on the currently estimated probability values, its value changes with each iteration. Therefore, if the model for additional sensor also depends on the currently estimated probability values, we must ensure that its values change at the same rate as those of the model for the visual sensors, otherwise one model or the other consistently prevails.

To avoid this issue we propose a model that is independent of the currently estimated probabilities, and show that it can influence the POM's convergence only in its early iteration steps. Such an auxiliary sensor can be used to improve the robustness of the system when the information from visual sensors is ambiguous and would cause the algorithm to converge to an incorrect solution.

The proposed approach has been experimentally verified on a problem of person localization in a cluttered environment using a radio tracker and camera-based POM. We demonstrated the strengths and weaknesses of both systems when used separately, and showed that by fusing them together, improved localization results are obtained, both in terms of



(a) The complete error CDF

(b) The error CDF for the inliers only (threshold: 1 *m*)

Fig. 2. The error cumulative density function for the Ubisense radio tracking system, camera-based POM and their fusion.

2-D error and in terms of false positive and false negative detections.

Acknowledgment

This work was supported by the research program P2-0095 and the research grant 1000-10-310118, both by the Slovenian Research Agency.

References

[1] J. Hightower and G. Borriello, "Location systems for ubiquitous computing," *Computer*, vol. 34, no. 8, pp. 57–66, Aug. 2001.

[2] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 6, pp. 1067–1080, 2007.

[3] J.-S. Franco and E. Boyer, "Fusion of multi-view silhouette cues using a space occupancy grid," *Computer Vision, IEEE International Conference on*, vol. 2, pp. 1747–1753, 2005.

[4] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 2, pp. 267–282, 2008.

[5] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, December 2006.

[6] C. Santiago, A. Sousa, M. Estriga, L. Reis, and M. Lames, "Survey on team tracking techniques applied to sports," in *Autonomous and Intelligent Systems (AIS), 2010 International Conference on*, 2010, pp. 1–6.

[7] A. Mittal and L. S. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene," *International Journal of Computer Vision*, vol. 51, pp. 189–203, 2003.

[8] J. Kang, I. Cohen, and G. Medioni, "Tracking people in crowded scenes across multiple cameras," in *Asian conference on computer vision*, vol. 7, 2004, p. 15.

[9] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for easyliving," *Visual Surveillance, IEEE Workshop on*, vol. 0, p. 3, 2000.

[10] D. Beymer, "Person counting using stereo," *Human Motion, Workshop on*, vol. 0, pp. 127–133, 2000.

[11] D. Yang, H. González-Baños, and L. Guibas, "Counting people in crowds with a real-time network of simple image sensors," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 122–129 vol.1.

[12] D. Delannay, N. Danhier, and C. De Vleeschouwer, "Detection and recognition of sports(wo)men from multiple views," in *Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference on*, 2009, pp. 1–7.

[13] R. Muñoz-Salinas, R. Medina-Carnicer, F. Madrid-Cuevas, and A. Carmona-Poyato, "People detection and tracking with multiple stereo cameras using particle filters," *Journal of Visual Communication and Image Representation*, vol. 20, no. 5, pp. 339–350, 2009.

[14] J. Berclaz, F. Fleuret, and P. Fua, "Principled detection-by-classification from multiple views," in *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, vol. 2, January 2008, pp. 375–382.

[15] A. Alahi, Y. Boursier, L. Jacques, and P. Vandergheynst, "Sport players detection and tracking with a mixed network of planar and omnidirectional cameras," in *Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference on*, 2009, pp. 1–8.

[16] T.-Y. Lee, T.-Y. Lin, S.-H. Huang, S.-H. Lai, and S.-C. Hung, "People localization in a camera network combining background subtraction and scene-aware human detection," in *Proceedings of the 17th international conference on Advances in multimedia modeling - Volume Part I*, ser. MMM'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 151–160.

[17] M. Anne, J. L. Crowley, V. Devin, and G. Privat, "Localisation intra-bâtiment multi-technologies: Rfid, wifi et vision," in *Proceedings of the 2nd French-speaking conference on Mobility and ubiquity computing*, ser. UbiMob '05. New York, NY, USA: ACM, 2005, pp. 29–35.

[18] A. Cattoni, A. Dore, and C. Regazzoni, "Video-radio fusion approach for target tracking in smart spaces," in *Information Fusion, 2007 10th International Conference on*, 2007, pp. 1–7.

[19] L. Marchesotti, R. Singh, and C. Regazzoni, "Extraction of aligned video and radio information for identity and location estimation in surveillance systems," in *Proc. of the 7th International Conference on Information Fusion*, 2004, pp. 316–321.

[20] C. Town, "Fusion of visual and ultrasonic information for environmental modelling," in *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, 2004, p. 124.

[21] S. Shivappa, M. Trivedi, and B. Rao, "Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692–1715, Oct. 2010.

[22] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 734–741 vol.2.

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, 2005, pp. 886–893 vol. 1.

[24] "Ubisense RTLS - Ubisense," http://www.ubisense.net/en/.

[25] S. Gezici, Z. Tian, G. Giannakis, H. Kobayashi, A. Molisch, H. Poor, and Z. Sahinoglu, "Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks," *Signal Processing Magazine, IEEE*, vol. 22, no. 4, pp. 70 – 84, 2005.

[26] J.-Y. Bouguet, "Camera Calibration Toolbox for Matlab," http://www.vision.caltech.edu/bouguetj/calib_doc/index.html.

[27] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *Proceedings of the eleventh ACM international conference on Multimedia*, ser. MULTIMEDIA '03. New York, NY, USA: ACM, 2003, pp. 2–10.

[28] "POM: Probabilistic Occupancy Map," http://cvlab.epfl.ch/software/pom/.