Multi-camera and radio fusion for person localization in a cluttered environment

Rok Mandeljc, Janez Perš, Matej Kristan, Stanislav Kovačič

Faculty of Electrical Engineering, University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenia
rok.mandeljc@fe.uni-lj.si

Abstract
We investigate the problem of person localization in a cluttered environment. We evaluate the performance of an Ultra-Wideband radio localization system and a multi-camera system based on the Probabilistic Occupancy Map algorithm. After demonstrating the strengths and weaknesses of both systems, we improve the localization results by fusing both the radio and the visual information within the Probabilistic Occupancy Map framework. This is done by treating the radio modality as an additional independent sensory input that contributes to a given cell's occupancy likelihood.

1. Introduction

In the past decade, the object localization and tracking has been an active research topic. Various localization solutions based on different sensor modalities have been proposed [10, 12]. The interest in the object localization and tracking coincides with the rise of applications requiring such information, which have seen adoption in fields such as manufacturing, military, surveillance and security, transport and logistics, childcare and tracking in sports. Two of the most prominent areas of the research are tracking using video cameras and radio-based localization.

In this paper, we investigate the problem of person localization in a cluttered environment, which poses a challenge both for camera and radio based approaches. On one hand, objects such as the office furniture occlude portions of the subjects' bodies, causing difficulties in accurate and reliable camera-based position estimation. On the other hand, the presence of radio-reflective metallic surfaces, in conjunction with obstacles, leads to multipath-related problems in the radio-based localization.

We propose a fusion scheme for combining the information from two such localization systems to overcome their problems in a cluttered environment. The first system is based on the Ultra-Wideband (UWB) radio technology and the second is a multi-camera-based system using the Probabilistic Occupancy Map (POM) algorithm [8, 7]. We demonstrate the strengths and weaknesses of both systems and verify that the fusion of their information within the POM improves the localization results.

2. Related work

Object detection and tracking using video cameras has a long tradition, with many different approaches having been proposed [17]. Recently, multi-camera approaches have gained prominence due to the ability to handle complex occlusions and difficult scenes [16, 8, 7, 14].

To our knowledge, little has been done in the field of fusing the visual information with the information from the UWB radio location systems. There have, however, been various attempts at the person localization via fusion of the computer vision and other radio techniques, such as RFID [3] or WiFi signal strength [3, 5, 13]. Town [15] describes a sentient office, in which the visual information from calibrated cameras is combined with the location events from an ultrasonic tracking system. Similarly, Checka *et al.* [6] provide a probabilistic framework for fusion of a stereo-based visual foreground detection and an audio-based localization for detecting and tracking people in a room.

Most of the above approaches are concerned with tracking, which involves temporal filtering. In contrast, in [8], authors focus on a robust and reliable frame-by-frame detection. Later, in [7], they use the proposed algorithm in conjunction with the dynamic programming to perform tracking. Similarly to their approach, it is our goal to first achieve a reliable (frame-by-frame) detection and localization, which can later be incorporated into a tracking framework.

3. Localization

Two of the most prominent and widely-used indoor localization approaches are localization using (multiple) video cameras and radio-based localization. Based on two different technologies, each approach has its advantages and disadvantages and both have difficulties when used in a cluttered environment. We consider an environment cluttered when it contains objects such as office furniture and equipment (desks and chairs, monitors, printers, computer cases, etc.). These (partially) occlude the individuals we wish to localize, thus causing difficulties in accurate and reliable camera-based position estimation. On the other hand, radio-reflective (metallic) surfaces in conjunction with obstacles lead to multipath-related problems in the radio-based localization, decreasing its performance as well.

3.1. Ubisense Real-Time Localization System

The radio localization system we use is the Ubisense Real-Time Localization System [2], which is based on the Ultra-Wideband (UWB) radio technology [9]. The system comprises a network of time-synchronized sensors (receivers) and tags (emitters) placed on the objects to be tracked. Two types of tags are available; the *slim tags* (directional emitters) for tagging the personnel and the *compact tags* (omnidirectional emitters) for tagging the equipment.

The tag's position is determined using the Time Difference of Arrival (TDoA) and the Angle of Arrival (AoA) measurements that are combined using a least-squares algorithm [12]. The tags' position information is, along with the estimations of a tag position's standard error, available via the software platform. The advertised accuracy of the system is 15 *cm*, with 99% of errors being within 30 *cm*; however, our preliminary experiments indicate that performance in a cluttered environment is actually lower due to obstacles and signal reflections.

3.2. Probabilistic Occupancy Map

For the camera-based localization, we employ multiple cameras and the POM algorithm [8, 7]. The use of an occupancy map involves discretization of the area of interest into a grid and estimating the probability of a person being present in each of the cells. POM estimates the posterior probabilities based on the background subtraction images in each of the cameras and an appearance model that approximates the silhouettes of the individuals with a family of rectangles.

The algorithm uses a generative model that, based on the currently estimated probability field, produces a synthetic image for each of the cameras. The probability field is then iteratively optimized, so that the generated images match the input binary images, until convergence to a fixed point is reached. The reference implementation of the algorithm is openly available [1].

The accuracy of the algorithm is defined by the size of the grid cells. As noted in [8], failures with POM usually occur when the motion detection blobs are severely inaccurate (lack of a contrast between a person and the background) or when a person is visible in only one view. However, due to a simple appearance model, the algorithm should be able to cope even with low quality binary blobs.

3.3. Fusion

Fusion of the information from sensors of different modalities has been proposed before to overcome the difficulties of an individual sensor [10, 12]. In principle, we could perform the fusion by taking the output of the Ubisense and of the POM and fusing them using a Bayesian probabilistic framework. However, in cases when the POM detection fails, all visual information would be discarded.

Therefore, in order to prevent the POM from converging to an incorrect solution when the visual information is ambiguous, we fuse the position information obtained from the Ubisense directly into it. 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. Its output, the tags' position information, is converted to the cell occupancy information and then fused with the visual information within the POM.

It can be seen from the equation for the posterior probability of a person being present in the i-th cell, $P(X_i = 1|V)$ [8, Appendix A] that the integration of the information from the Ubisense system into the POM algorithm is possible by introducing an additional likelihood ratio term to the product in the denominator, as if we had an additional camera:

$$\frac{P(U|X_i = 0, X_{[i]})}{P(U|X_i = 1, X_{[i]})} \tag{1}$$

The likelihood $P(U|X_i = 0, X_{[i]})$ is modeled by a uniform distribution, evaluated at the center of each cell. U denotes the Ubisense information, similarly to V denoting visual information in the original equation. Due to the normalization across all the cells, the value for each cell is actually $1/N$, where N is the number of all cells. A similarly obtained uniform distribution is used for $P(U|X_i = 1, X_{[i]})$, and is mixed with a mixture of Gaussians that represents the Ubisense information. It is obtained from Ubisense tag locations and their estimated standard error, and evaluated at the center of each cell, followed by the normalization across all the cells. The obtained likelihood ratio (1) is thus low for the cells close to the Ubisense tag detections and high for the cells far from the Ubisense tag detections. A low ratio increases, while a high ratio decreases the posterior probability of presence in a cell. In the case when no information is available from the Ubisense system (i.e. no tags in the room), the ratio equals to 1 and does not affect the posterior probability of the cell occupancy. The likelihoods are modeled as independent of the currently estimated values of the probability field $X_{[i]}$.

4. Experimental setup

For the experiment, four calibrated Axis 207W IP cameras and four Ubisense sensors were placed in the corners of a 7.1×6.9 *m* room, at the height of about 2 *m* (Figure 2). Three lines were marked on

the floor for the individuals to walk on; due to limits in cameras' field of view, most of the locations along the lines are covered by two cameras.

A 6-minute sequence involving three individuals wearing the Ubisense slim tags was captured. Data was processed off-line; video from the cameras (640×480 at 25 FPS) was streamed to files and later synchronized with the captured Ubisense position data using timestamps. The was processed using the lens distortion correction [4], followed by the background subtraction algorithm [11] implemented in the OpenCV library (using the default parameters). 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 thus fits into multiple cells; however, multiple adjacent detections could be merged during the post-processing.

For the experiment, the reference POM algorithm was modified to integrate the Ubisense position data. 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. We consider a cell to be occupied when the posterior probability of presence $P(X_i = 1|V)$ as estimated by the POM is greater or equal to 0.5.

For the performance evaluation, every 25th frame was taken and manually annotated; the ground truth data was obtained by clicking on the persons' heads in at least two views, thus obtaining their 2-D position. This way, 383 frames were annotated with the ground truth positions. The error of the ground truth data was calculated from the distances between an obtained ground truth point and its projection to the corresponding line on the floor. The mean and standard deviation of error were found to be 3.6 *cm* and 2.8 *cm*, with 95% and 99% of errors lying within 8.8 and 12.5 *cm* respectively.

5. Performance metric

The performance metric used for the evaluation of a system's performance is the 2-D position error — the Euclidean distance between the estimated and the ground truth 2-D position. The error cumulative density function (CDF) is used to obtain the error boundaries that cover 95% or 99% of all errors, which is our metric for the system precision. However, evaluating just the 2-D position error does not account for the cases when the system completely fails to detect a person.

In the case of background subtraction-based POM algorithm, the identity of a person occupying a cell is unknown and there is no way of telling to which ground truth point does the given detection (occupied cell) belong. Therefore, we assume that each detection belongs to the closest ground truth point and vice versa. A threshold 1 *m* is used to decide whether a detection is an inlier or an outlier. If the distance between a detection and the closest ground truth point is greater than the threshold, the detection has no corresponding ground truth point and is called *a phantom*. Similarly, if the distance from a given ground truth point to the closest detection is greater than the threshold, the ground truth point has no corresponding detection — we are dealing with *a missing detection*. When both a phantom and a missing detection are present, we assume they correspond to each other and call such a case *an inaccurate detection*. When evaluating a system's performance, the number of the occurrences of each error type is considered in addition to the 2-D position error.

With the Ubisense, the identity is explicitly given; a tag can be detected only if it is present in a room, and therefore there are no phantom detections[1]. A missing detection occurs when the signal from a tag

[1]In practice, the UWB signal penetrates through walls, therefore a phantom detection can occur when a tag is present in the adjoining room; however, we ignore such cases because they are not equivalent to the phantom detections in POM.

cannot reach the receivers and an inaccurate detection occurs when the signal arrives to the receivers via reflections with the direct path being blocked. For comparison with POM, we classify a Ubisense detection as an inaccurate if its distance to the corresponding ground truth point exceeds 1 m.

6. Results and discussion

As can be seen from the summarized results in Table 1a and the plots in Figure 1a, the error CDF curve for camera-based POM reaches the 95% mark sooner than the Ubisense's error CDF curve. With the 99% error boundary, however, the situation is reversed. This can be explained by the gross errors in camera-based POM detections — the phantoms and the inaccurate detections. If those outliers (the detections with error greater than 1 m) are discarded, the error CDF for POM reaches both marks sooner than for the Ubisense (Figure 1b). Therefore we can conclude that the Ubisense system has lower position accuracy than the localization with camera-based POM, however it is more reliable in terms of the gross errors. Conversely, the camera-based POM offers higher accuracy, but at the price of the reduced reliability.

By fusing the Ubisense data into the POM, most of the phantom and inaccurate detections are resolved, which results in a much steeper error CDF curve in Figure 1a. The error CDF curve for the inliers only (Figure 1b) shows only a marginal improvement compared to the camera-based POM. 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 does not impair the correct detections, even in the cases when the Ubisense position estimation is off (Figure 2a). Examples of a phantom and an inaccurate detection and their resolution are shown in Figure 2a and Figure 2b respectively.

Table 1b summarizes the number of the phantom, missing and inaccurate detections. Again, it can be seen that most of the phantom and inaccurate detections from camera-based POM are resolved by the proposed fusion method. The number of the missing detections is also slightly lowered. However, the missing detections occur in cases of almost complete absence of the visual information (occlusions or poor background subtraction) and cannot be corrected even with the additional information from the Ubisense (Figure 2c). Nor should it, because otherwise the fusion would introduce new phantoms in cases when the Ubisense position estimation is off, such as the one shown in Figure 2a.

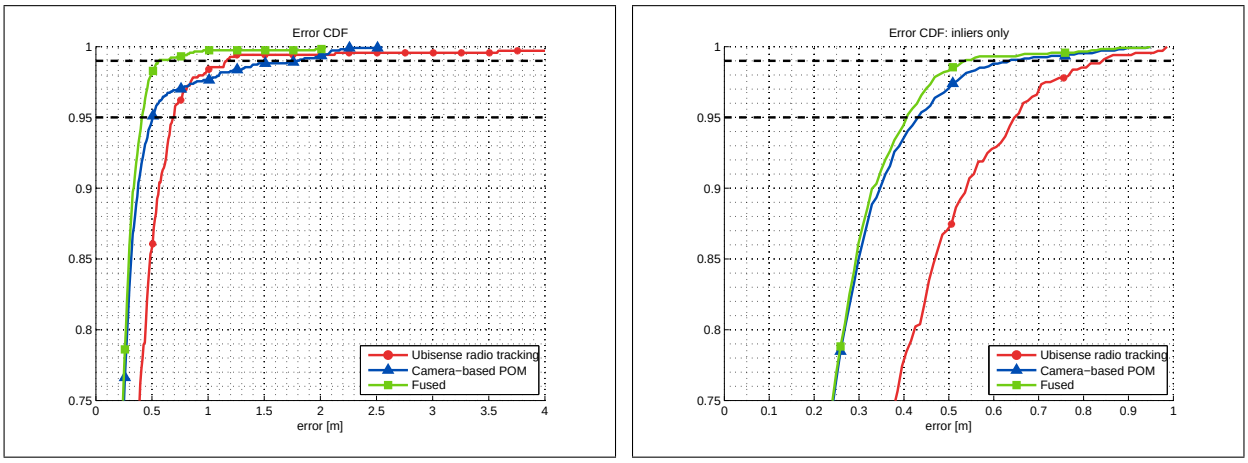
	Ubisense	**POM**	**fused**
mean [m]	0.32	0.23	0.19
std [m]	0.34	0.26	0.15
error boundaries			
95% [m]	0.70	0.51	0.41
99% [m]	1.17	1.79	0.66
error boundaries (inliers only)			
95% [m]	0.64	0.43	0.41
99% [m]	0.84	0.64	0.57

(a) Error statistics

	Ubisense	**POM**	**fused**
phantom	N/A	13	2
missing	9	62	42
inaccurate	11	13	1

(b) Phantoms, missing and inaccurate detections

Table 1. The results summary for the Ubisense system, camera-based POM and their fusion.



(a) The complete error CDF

(b) The error CDF for the inliers only

Figure 1. The error cumulative density function for the Ubisense system, camera-based POM and their fusion.

7. Conclusion

We have proposed a novel fusion scheme for combining the information from a camera-based and a radio-based localization system using the POM algorithm in order to improve localization in a cluttered environment. The novelty of our approach lies in the direct fusion of both visual and radio information within the POM framework, showing its potential as a general multi-modal sensor fusion framework.

We analyzed the performance of the UWB-radio-based Ubisense localization system and the system comprising multiple cameras and the POM algorithm. The results indicate that the Ubisense system has lower accuracy than the localization with camera-based POM, however it is more reliable in terms of the gross errors, while the camera-based POM offers higher accuracy, but at the price of the reduced reliability. The proposed fusion method that integrates the Ubisense position data directly into the POM algorithm resolves the gross errors that occurred with the camera-based system while retaining its accuracy. The proposed method does not solve the problem of the missing detections caused by poor background subtraction results; however, this should be addressed by other means, such as relaxing the background subtraction algorithm parameters or integrating additional visual cues (e.g. people detectors).

At this point we did not perform any in-depth analysis of the data model for the Ubisense measurements and the parameters for their integration into the POM. In the future, we plan to perform a formal mathematical treatment of the model in order to obtain the parameters for the optimal fusion and verify its results using a larger data set in terms of the annotated ground truth data.

References

[1] POM: Probabilistic Occupancy Map. `http://cvlab.epfl.ch/software/pom/`.

[2] Ubisense RTLS - Ubisense. `http://www.ubisense.net/en/`.

[3] M. Anne, J. L Crowley, V. Devin, and G. Privat. Localisation intra-bâtiment multi-technologies: RFID, wifi et vision. *UbiMob*, 5:29–35, 2005.

[4] Jean-Yves Bouguet. Camera Calibration Toolbox for Matlab. `http://www.vision.caltech.edu/bouguetj/calib_doc/index.html`.

[5] A. F. Cattoni, A. Dore, and C. S. Regazzoni. Video-radio fusion approach for target tracking in smart spaces. In *Proceedings of the 10th International Conference on Information Fusion*, page 1–7, 2007.

[6] N. Checka, K. Wilson, V. Rangarajan, and T. Darrell. A probabilistic framework for multi-modal multi-person tracking. *Computer Vision and Pattern Recognition Workshop*, 9:100, 2003.

[7] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, 2007.

[8] F. Fleuret, R. Lengagne, and P. Fua. Fixed point probability field for complex occlusion handling. Technical Report IC/2004/87, EPFL, October 2004.

[9] S. Gezici, Z. Tian, G. B. Giannakis, H. Kobayashi, A. F. Molisch, H. V. Poor, and Z. Sahinoglu. Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks. *IEEE Signal Processing Magazine*, 22(4):70–84, 2005.

[10] J. Hightower and G. Borriello. Location systems for ubiquitous computing. *Computer*, 34(8):57–66, August 2001.

[11] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian. Foreground object detection from videos containing complex background. In *Proceedings of the 11th ACM International Conference on Multimedia*, page 10, 2003.

[12] H. Liu, H. Darabi, P. Banerjee, and J. Liu. Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 37(6):1067–1080, 2007.

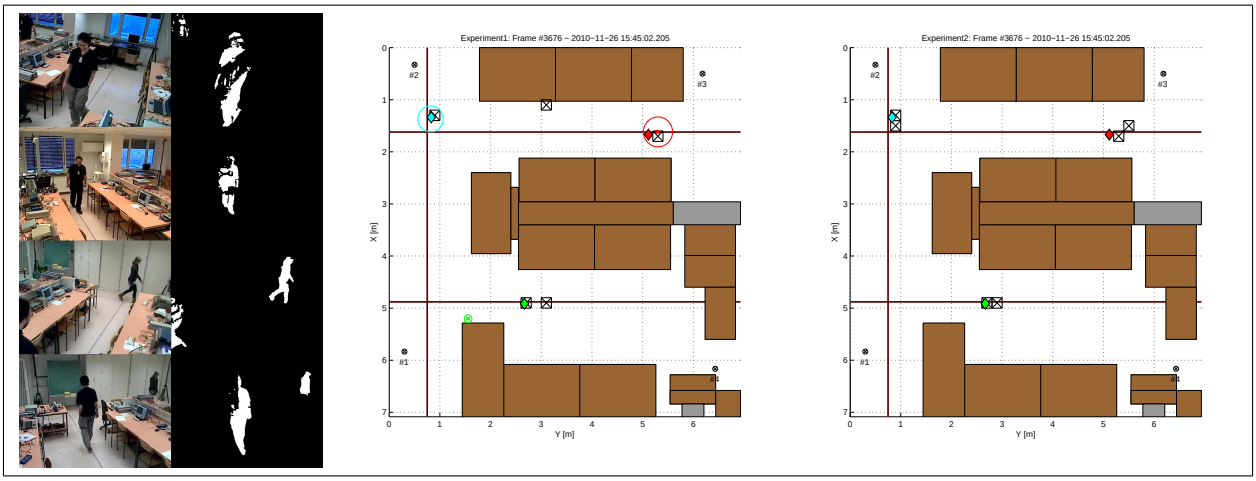
[13] L. Marchesotti, R. Singh, and C. S. Regazzoni. Extraction of aligned video and radio information for identity and location estimation in surveillance systems. In *Proceedings of International Conference of Information Fusion*, 2004.

[14] R. Muñoz-Salinas, R. Medina-Carnicer, F. J. Madrid-Cuevas, and A. Carmona-Poyato. People detection and tracking with multiple stereo cameras using particle filters. *Journal of Visual Communication and Image Representation*, 20(5):339–350, 2009.

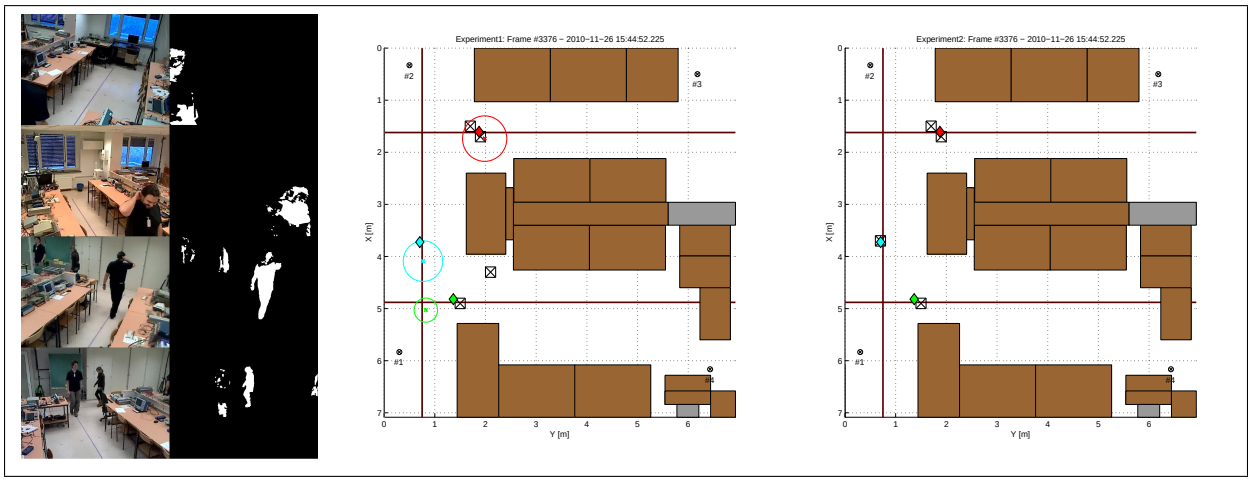
[15] C. Town. Fusion of visual and ultrasonic information for environmental modelling. In *Proceedings of Conference on Computer Vision and Pattern Recognition Workshop*, page 124, 2004.

[16] D. B Yang, H. H González-Baños, and L. J Guibas. Counting people in crowds with a real-time network of simple image sensors. In *Proceedings of the 9th IEEE International Conference on Computer Vision.*, page 122–129, 2003.

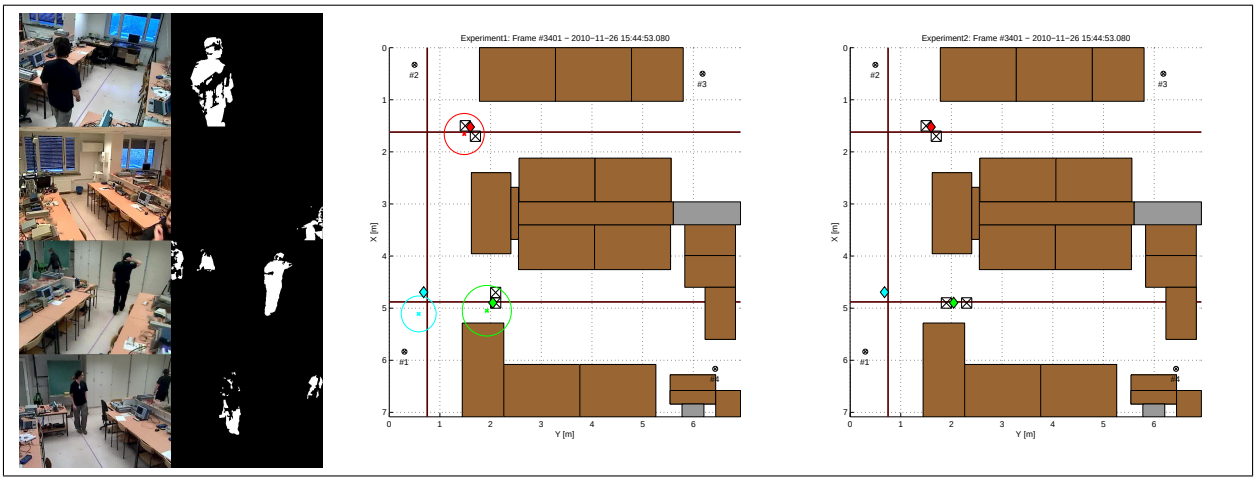
[17] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4), December 2006.



(a) A phantom detection and its resolution; the inaccuracy of a position given by the Ubisense does not introduce new phantoms.



(b) An incorrect detection and its resolution



(c) A case when a missing detection cannot be resolved

Figure 2. Examples of localization. On the left, the video frames and the corresponding background subtraction results are shown. On both top-view plots, the ground truth positions are marked with the colored diamonds. The left top-view plot displays the results of the Ubisense system (colored crosses denote the position and colored circles the corresponding estimated standard error) and of the POM (the black crossed rectangles indicate the occupied cells). The right top-view plot displays the results of the fusion (black crossed rectangles).