

# A Semantic World Model for Urban Search and Rescue Based on Heterogeneous Sensors

Johannes Meyer<sup>2</sup>, Paul Schnitzspan<sup>1</sup>, Stefan Kohlbrecher<sup>1</sup>, Karen Petersen<sup>1</sup>,  
Mykhaylo Andriluka<sup>1</sup>, Oliver Schwahn<sup>1</sup>, Uwe Klingauf<sup>2</sup>, Stefan Roth<sup>1</sup>,  
Bernt Schiele<sup>1,3</sup>, and Oskar von Stryk<sup>1</sup>

<sup>1</sup>Department of Computer Science, TU Darmstadt, Germany

<sup>2</sup>Department of Mechanical Engineering, TU Darmstadt, Germany

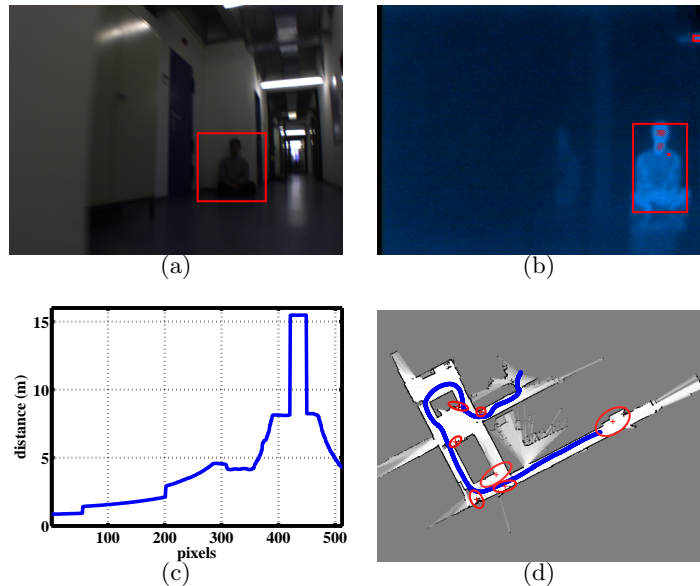
<sup>3</sup>MPI Informatics, Saarbrücken, Germany

**Abstract.** In urban search and rescue scenarios, typical applications of robots include autonomous exploration of possibly dangerous sites, and the recognition of victims and other objects of interest. In complex scenarios, relying on only one type of sensor is often misleading, while using complementary sensors frequently helps improving the performance. To that end, we propose a probabilistic world model that leverages information from heterogeneous sensors and integrates semantic attributes. This method of reasoning about complementary information is shown to be advantageous, yielding increased reliability compared to considering all sensors separately. We report results from several experiments with a wheeled USAR robot in a complex indoor scenario. The robot is able to learn an accurate map, and to detect real persons and signs of hazardous materials based on inertial sensing, odometry, a laser range finder, visual detection, and thermal imaging. The results show that combining heterogeneous sensor information increases the detection performance, and that semantic attributes can be successfully integrated into the world model.

## 1 Introduction

Modeling the world in complex environments is a crucial aspect on the way toward reliable, intelligent, and autonomous search and rescue robots. As motivated in [1–3], it is desirable not only to infer a geometrically interpretable map, but also to integrate semantic attributes to enable high-level scene interpretation. In urban search and rescue (USAR), reliable robots have to provide a semantically meaningful interpretation of objects within a scene (e.g. victims in collapsed buildings) [4]. In unconstrained environments – as is the case in USAR scenarios – relying only on one type of sensor is often insufficient, while fusing complementary information (i.e. information from different types of sensors) enables semantic interpretability of scenes and superior reliability.

The main goal of this paper is to propose a mobile robot system for autonomous detection of victims in USAR scenarios. The system is capable of autonomous navigation and map learning, and localizes victims and objects of interest in a 3D world coordinate system. Our setting approaches actual search and rescue operations in realism and complexity: Real human victims have to be localized in unstructured environments, even in the presence of background clutter and multiple thermal



**Fig. 1.** Examples of sensor and world model data: (a) Visual image with victim detection. (b) Thermal image with heat detections. (c) Range samples along the horizontal axis of the image. (d) Snapshot of the semantic world model with estimated victim locations denoted by the red covariance ellipses.

sources such as office equipment and heating. To this end, we expand the scenario beyond the current RoboCup Rescue competition, in which the environment is built of simple structural elements and a thermal camera is sufficient for victim localization in most cases. By adding objects of interest to the metric map of the environment, we augment the robot’s environment with semantic information, which can then be utilized for decision making by human operators. Our system is able to achieve high performance even for cluttered and complex datasets (see Fig. 1). All information is processed onboard and in real-time, as is crucial for realistic rescue deployments.

In our system visual information is supplied to a generic object detector that allows detecting structured objects and assigns them a semantic meaning (e.g. upper bodies of victims or hazardous material signs). To avoid relying on a single source of information, we consider the information from all sensors simultaneously, and derive a generic model that is able to leverage complementary information. As motivated in [5], merging different sources of information helps achieving higher levels of performance in victim detection. However, to the best of our knowledge, this work is the first attempt at integrating and evaluating state-of-the-art visual object detection with an autonomous USAR robot system.

Sec. 2 describes our system, while in Sec. 3 the sensors and detection algorithms for victim and object detection are introduced. Afterwards two approaches for sensor fusion are presented in Sec. 4. Experimental validation is presented in Sec. 5.

**Related work.** Enhancing geometric world models with semantic information is motivated in [6], where an indoor environment is described by corridors, doors, and types of rooms. Features are extracted from camera images and laser range

data, and subsequently classified via hidden Markov models. The paper shows that semantic world models are helpful for behavior planning. In our context, we focus on detecting objects of interest, for example victims, rather than rooms and doors.

State-of-the-art algorithms for people detection frequently rely on only one type of sensor. In particular, a large body of literature exists for people detection in visual camera images and video sequences. Much of the progress in this area has been achieved by combining statistical machine learning techniques with robust image representations. High variability in people poses and appearance is often addressed using part-based models, in which parts either correspond to anatomically meaningful body parts [7], or are automatically inferred from data [8, 9]. In [10] an image descriptor based on histograms of oriented gradients (HOG) is proposed, which is combined with discriminative SVM classifiers in order to exhaustively scan an image for people hypotheses. This representation has been further extended in [11] to incorporate motion information and has been used as basis for multiple approaches to people and object detection [12–14]. Although being conceptually simple, HOG-based detectors belong to the best people detection methods published to date [15].

Thermal images can be used to detect upright people [16, 17], to distinguish upright people from people lying on the ground [18], and to detect body parts and arbitrary postures of humans [19]. Laser range finders are frequently used to detect upright standing or walking people, either by tracking the upper body [20], the legs [21], or both [22]. All of these algorithms have restrictions and work only in specific situations. Some require upright standing or walking people, others assume to have no other heat sources than humans in the sensing range. These drawbacks can be overcome by using several heterogeneous, complementary types of sensors. However, combinations of several heterogeneous sensors have been shown to perform better than each classifier alone. In [23, 24] detections from a laser range finder are used to classify regions of interest for a visual detector, and in [5] several independent classifiers relying on heat, skin color, motion and face detection are fused using Markov random fields.

## 2 System Overview

Since the locations of victims need to be specified in world coordinates, the robot pose and a metric map have to be estimated using a combination of inertial sensing and simultaneous localization and mapping (SLAM). These estimates are continuously updated over time and used for integrating victim hypotheses obtained from different sensor types. Simultaneously with the pose and map estimation at each time step, our system generates a set of object hypotheses using a visual object detector and thermal-camera based detector, which are used as a basis for sensor fusion.

We explore two complementary approaches for sensor fusion. In the first, which we denote as explicit sensor fusion, we integrate information from different sensors directly in the sensor space using known transformations between different sensor modalities (i.e. the mapping between thermal and visual images). This allows to use known dependencies of sensor signals to either amplify or attenuate the confidence in the measurements.

In the second approach, which we denote as implicit sensor fusion, the global belief is updated independently for each observation. The advantage of using complementary sensors is realized through accumulation of positive evidence in the world model. Integrating hypotheses into the model requires an association step for matching a hypothesis to an already known object. The case in which previously unknown objects have been found must be considered separately. Once association is established, the matching is taken for granted and the corresponding victim location estimate and evidence is updated by using an extended Kalman filter (EKF). Integration of observations into the global belief state can therefore be considered to be a method for temporal sensor fusion. Additionally, confidence in a hypothesis is influenced by negative observations, where the absence of expected detections or contradictory measurements reduce the confidence value.

When applying implicit sensor fusion the observations from different sensors are integrated independently into the world map, while explicit sensor fusion is an optional step that precedes the integration step, and is primarily used to increase the reliability of observations.

## 2.1 World Model

World models generally account for a mathematical description of the environment, with different aspects being considered important depending on the application. In our USAR scenario the model is formed by a representation of building geometry and additional semantic information, like the location of people and objects of interest. By applying additional high-level knowledge to the model, it can be easily enriched with more detailed information in future work, e.g. classification of places, a graph of passable paths through a building, or estimates of hazardousness of specific locations. Based on this high-level description of the environment, the robot is able to plan reasonable future actions and – when integrating human operators – is able to deliver valuable information to rescue teams, e.g. to guide them to detected victims.

The robot state vector  $\mathbf{y}_k$  contains the estimated 6DOF robot pose as well as translational and angular velocities in the global coordinate system and is updated at discrete timesteps  $t = t_k$ . The location of objects, including the victims, is referred to as  $\mathbf{x}_k^j$  with  $j$  being an index variable over the estimates. The objects are modeled as points, ignoring their spatial extent. In this paper we assume the world to be static apart from the movement of the robot itself. The number of objects is not known in advance. Besides the location information we introduce the probability  $\pi_k^j$  that object  $j$  is detected correctly as a measure of confidence, which is incrementally updated with each new sensor reading and typically increases when more detections of the same object occur.

The process of world modeling requires inference in state space from measurements given in sensor-space. Since sensors are error-prone, a probabilistic model description is used here. We choose a Gaussian representation for the continuous state variables, with estimated means  $\hat{\mathbf{y}}_k$  and  $\hat{\mathbf{x}}_k^j$ , and variances  $C_k$  and  $P_k^j$ , respectively.

## 2.2 Simultaneous localization and mapping (SLAM)

State estimation of the vehicle and a map is performed by two components. A 2D pose and map estimate is provided by a module using incremental maximum likelihood alignment of laser scans with the estimated map. The map is represented

by a discrete grid and updated using the log-odd probabilities of occupancy [25, 26].

Estimation of the robot state  $\mathbf{y}_k$  is performed by an extended Kalman filter (EKF) integrating observations from all available sensors. Attitude estimation is provided by a built-in IMU and compass, while position estimation is provided by wheel encoders and the 2D pose estimation updates from the SLAM module. For the USAR scenarios described in this work, our approach is sufficiently accurate as to not require multiple map hypotheses (e.g. using a Rao-Blackwellized particle filter), or explicit loop closure.

### 3 Victim and object detection

In order to enrich the map with semantic information, we perform on-board detection of objects of interest, which in our case correspond to people and dangerous materials marked with hazmat signs. In this paper we focus on the detection of upper bodies of people, since this allows to detect both standing people as well as possibly injured people sitting on the ground (see Fig. 1), and leave more complex cases for future work. Due to background clutter, partial occlusions and complex articulations, visual people detection is a difficult problem even in this somewhat restricted setting. In particular, state-of-the-art computer vision methods are still severely challenged by this task [12].

**Object detection.** In order to find initial hypotheses of people and hazmat signs in camera images, we use the popular sliding window approach. In this approach every image is exhaustively scanned over a range of positions and scales; for each position and scale a discriminative SVM classifier is used to make binary decisions about the presence or absence of an object. While seemingly expensive, the sliding-window approach is especially suitable for parallel implementation, since each object location can be examined independently of the rest of the image.

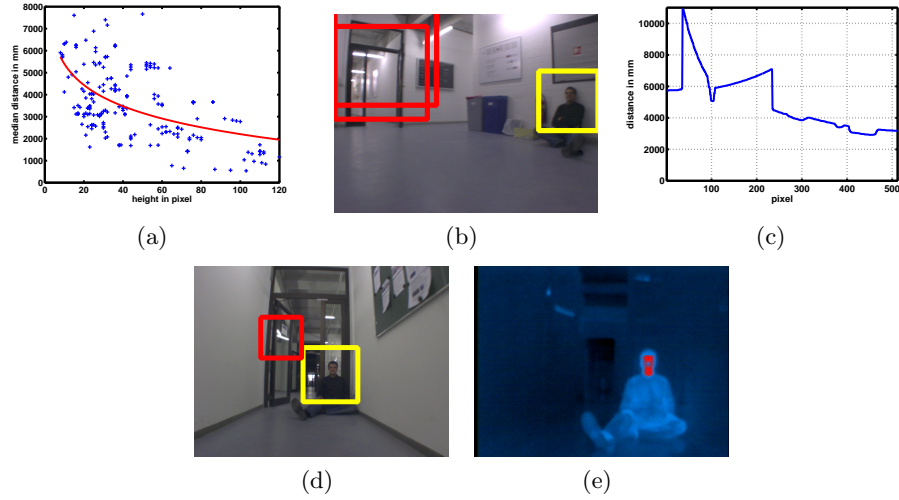
In order to describe the contents of the image at each particular location, we leverage recent results in computer vision and rely on a histogram of oriented gradients (HOG) descriptor [10]. In our system we scan the image with steps of 8 pixels and relative scale factors of 1.05. We use the GPU implementation [14] developed in our group, which allows to achieve real-time performance without sacrificing recognition performance.

The confidence  $s_k^{\text{vis}} \in [0, 1]$  of a hypothesis is calculated via a sigmoidal mapping:

$$s_k^{\text{vis}} = \frac{1}{1 + \exp(a \cdot f_k + b)}, \quad (1)$$

where  $f_k$  is the SVM score of the hypothesis, and  $a$  and  $b$  are parameters that are estimated by cross-validation [27].

**Object classification.** A HOG descriptor is especially well suited for capturing the characteristic shape of an object. However, it has shortcomings when it is necessary to distinguish between objects with similar shape, such as different hazmat signs, all of which have a rhombus shape and differ mainly in color, internal patterns and text. In order to identify hazmat signs we augment the HOG descriptors with color histograms. For each hazmat sign hypothesis of the HOG based detector, we compute a color histogram in LAB color space and use this to



**Fig. 2.** (a) Correspondence between annotation height and distance. (b) Example detections of frame 561. (c) Scanline of frame 561. (d) Example detections of frame 287. (e) Thermal image of frame 287.

perform the final classification of hazmat signs by applying a k-nearest neighbor approach in combination with the  $\chi^2$ -distance. As our experiments demonstrate, the combination of HOG and color histograms yields good performance for hazmat sign classification (Sec. 5).

**Thermal victim detection.** In addition to victim hypotheses from visible light camera images, our system also creates a set of hypotheses based on images from a thermal camera. These thermal hypotheses are generated with a simple procedure that searches the images for large enough groups of connected pixels with temperature values within the human body temperature range. Each such group of pixels is used to generate a hypothesis. Although hypotheses generated by the thermal camera alone are significantly less reliable compared to hypotheses produced by the visual object detector, we found them to be effective in reducing the number of false positives.

We define a simple model for the confidence  $s_k^{\text{therm}} \in [0, 1]$  by counting the number of pixels within the person’s bounding box that have a temperature close to human bodies. This model is robust to small offsets in corresponding locations in visual and thermal images, which arise due to imprecise synchronization between these modalities.

## 4 Sensor fusion

The reliability of the entire victim detection framework can be increased by fusing victim hypotheses from different sensors and across time steps. Intuitively, the confidence of a detected victim should be increased if it is observed in several update steps or by different sensors and on the other hand stay below a certain threshold when it is only spotted once. We employ an extended Kalman filter (EKF) in order to update the locations  $\mathbf{x}_k^i$  of victims in our world model and integrate several

hypotheses across update steps. In parallel, the confidence  $\pi_k^j$  that the victim is present at the respective location is updated in a separate filter with the respective measurement confidence as described below.

For simplification of notation we assume without loss of generality that at most one hypothesis is observed in every update step  $k$ . We define the measurement  $\mathbf{z}_k$  to consist of the distance  $d_k$ , the bearing angle  $\alpha_k$ , and the relative vertical angle  $\beta_k$  between the hypotheses and the robot:

$$\mathbf{z}_k = [d_k \ \alpha_k \ \beta_k]^T = h(\mathbf{x}_k^j, \mathbf{y}_k) + \mathbf{v}_k, \quad (2)$$

where  $h(\cdot, \cdot)$  refers to a nonlinear measurement function that projects the victim's position into the world model.  $\mathbf{x}_k^j$  and  $\mathbf{y}_k$  denote the victim's position estimate and robot state vector respectively. The random vector  $\mathbf{v}_k$  is unbiased and uncorrelated Gaussian measurement noise with hand-tuned variance  $R$ . The measurement function also depends on the robot's state  $\mathbf{y}_k$ , which in turn is estimated with an EKF independently.

**Data association.** In order to find an optimal matching between measurements and existing estimates of victim locations we use the following probability of measurement  $\mathbf{z}_k$  given the index  $j$  and the position estimate  $\hat{\mathbf{x}}_{k-1}^j$  with variance  $P_{k-1}^j$ :

$$p(\mathbf{z}_k | j) \propto \mathcal{N}(\mathbf{z}_k; h(\hat{\mathbf{x}}_{k-1}^j, \hat{\mathbf{y}}_k), R + H_k^j P_{k-1}^j (H_k^j)^T) \quad (3)$$

with a first order approximation  $H_k^j$  of the measurement function  $h(\hat{\mathbf{x}}_{k-1}^j, \hat{\mathbf{y}}_k)$  at the current estimated means.

Whenever this probability  $p(\mathbf{z}_k | j)$  is above a previously defined threshold, we associate the new measurement to the best matching estimate with index  $j_k^* = \arg \max_j p(\mathbf{z}_k | j)$ . Otherwise, a new estimate is added to the world model as of a previously unobserved victim.

**Kalman filter updates.** We assume the victims to be static in our setting and therefore no explicit prediction step is needed. The measurement update equations of the Kalman filter are defined as:

$$K_k^j = P_{k-1}^j (H_k^j)^T \left( H_k^j P_{k-1}^j (H_k^j)^T + R \right)^{-1} \quad (4)$$

$$\hat{\mathbf{x}}_k^j = \hat{\mathbf{x}}_{k-1}^j + \lambda_k(\mathbf{z}_k, s_k) \cdot K_k^j \left( \mathbf{z}_k - h(\hat{\mathbf{x}}_{k-1}^j, \hat{\mathbf{y}}_k) \right) \quad (5)$$

$$P_k^j = \left( I - \lambda_k(\mathbf{z}_k, s_k) \cdot K_k^j H_k^j \right) P_{k-1}^j \quad (6)$$

where  $I$  denotes the identity matrix and  $s_k$  refers to the initial score of hypothesis  $k$  as will be explained below. The measurement update uses the confidence  $\lambda_k(\cdot, \cdot) \in [0, 1]$  of an observation as an additional factor to the gain matrix  $K_k^j$  to honor the observation quality and discard uncertain updates. The measurement confidence is of the form

$$\lambda_k(\mathbf{z}_k, s_k) = s_k \cdot \phi(\mathbf{z}_k, d^{\text{laser}}) \cdot \psi(\mathbf{z}_k), \quad (7)$$

where  $d^{\text{laser}}$  is the distance to the next obstacle measured with the laser scanner.  $\phi(\cdot, \cdot)$  imposes a prior on the estimated distance to  $\mathbf{z}_k$  and measured distance  $d^{\text{laser}}$  to the next obstacle and is proportional to a Gaussian with manually defined variance according to the uncertainty in the sensor measurements.  $\psi(\cdot)$  refers to a Gaussian height prior with mean 80cm (height of upper bodies) and manually

defined variance. By employing  $\phi(\cdot, \cdot)$ , we ensure that the size of an hypothesis approximately matches the size that we expect, and avoid false positives with inappropriate estimated and measured distances (see Fig. 2(b)).  $\psi(\cdot)$  guarantees that all objects appear at the expected height from the robot, while unlikely pitch angles are discarded.

**Label confidence update.** In the case that a new measurement is associated to a given estimate, we update the estimate’s label confidence according to the disjunctive combination of two binary random events, so that confidence is increased with every new measurement:

$$\pi_k^j = \pi_{k-1}^j + \lambda_k \cdot (1 - \pi_{k-1}^j). \quad (8)$$

If no measurement in time step  $k$  is available we decrease the label confidence of all victim estimates within the field of view by employing ”negative evidence”. Negative evidence is information that arises from the fact that the confidence of an estimate can decrease if it is not confirmed by sensor observations. Applying negative evidence to our algorithm helps to decrease the number of false alarms, as many false positives do not reoccur in consecutive time steps.

The negative update is applied to all objects  $j$  that should be visible in the image according to the current estimated map and positions, but have no detection event associated for the current time step. Their label confidence is reduced according to

$$\pi_k^j = \frac{p_{\text{miss}} \cdot \pi_{k-1}^j}{p_{\text{miss}} \cdot \pi_{k-1}^j + (1 - \pi_{k-1}^j)}. \quad (9)$$

The probability  $p_{\text{miss}}$  of missed detections is approximated as the inverse probability of the detector’s recall on the trained dataset.

**Implicit vs. explicit integration** We evaluate two different fusion schemes: implicit and explicit fusion. These two approaches differ in the way the complementary information of sensors is integrated. In our model this boils down to the treatment of the initial score  $s_k$  of an hypothesis in Eq. (7).

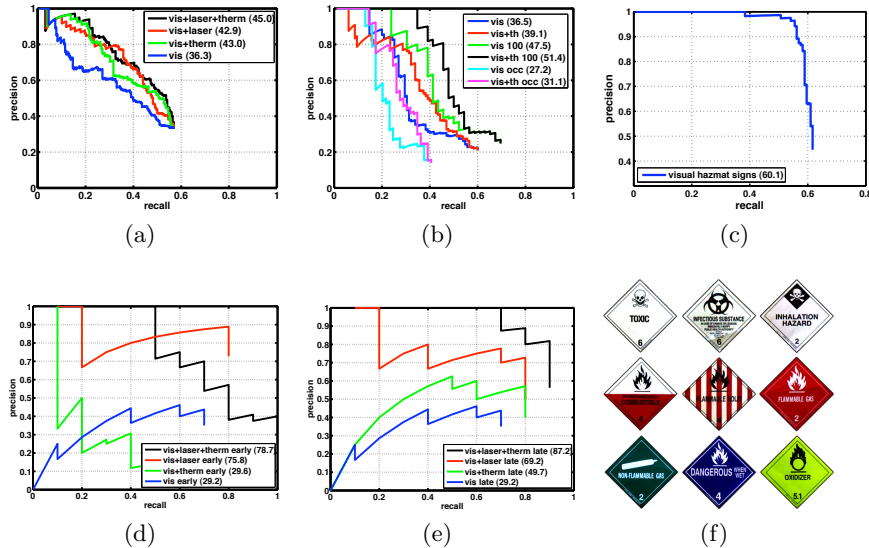
In the implicit fusion scheme we consider each hypothesis from both the visual light and thermal sensor as a new measurement that is either associated to a given estimate or enters the world model as a new estimate. In this setting the sensor fusion is implicitly handled with the Kalman filter, since the measurements of both sensors can be used for data association and updating the confidence. Here we directly use the visual or thermal score as initial score  $s_k$ :

$$s_k = \begin{cases} s_k^{\text{vis}}, & \text{if visual light hypothesis} \\ s_k^{\text{therm}}, & \text{if thermal hypothesis} \end{cases}. \quad (10)$$

In the explicit integration scheme we compute the overall detection confidence as a weighted sum from the individual scores of complementary sensors, yielding a single observation model where bearing and distance information is taken from the visible light bounding box only:

$$s_k = \gamma_1 \cdot s_k^{\text{vis}} + \gamma_2 \cdot s_k^{\text{therm}} + \gamma_3 \cdot s_k^{\text{laser}}, \quad (11)$$

where  $\sum_i \gamma_i = 1$  and the coefficients  $\gamma_i$  are trained with cross-validation. This additive formulation makes the model robust to sensor failures (e.g. due to partial



**Fig. 3.** (a, b) Single-frame people detection performance for different combination of sensors on “Hector Data 1” and “Hector Data 2” datasets. (c) Single-frame hazmat sign detection performance on “Hector Data 1” dataset. (d, e) People detection performance of the full system on the ”Hector Data 1” dataset for explicit and implicit sensor fusion schemes. (f) Collection of different hazmat signs.

occlusions) and takes relative importance of different sensors into account. The first two components of the mixture correspond to the probability of correct detection given the score of the SVM classifier and the output probability of the thermal victim detector for the same bounding box as defined in Sec. 3.

While more detailed integration of different sensor modalities is possible and we plan to explore it in the future, we opt for this re-estimation approach since it allows to decouple training of the visual object detector from the rest of the system, does not require exact synchronization between different sensor streams, and allows to use simple algorithms for integration of thermal and laser sensors.

In order to model  $s_k^{\text{laser}}$ , we fit a log-linear model to a set of jointly observed bounding boxes and laser range measurements as shown in Fig. 2(a).  $s_k^{\text{laser}}$  is set to a Gaussian computed at the difference between the predicted distance from the log-linear model and the median distance measured with the laser range finder. The variance is set by hand.

## 5 Experiments

We evaluate the performance of our system on the tasks of people and hazmat sign detection. In particular we quantify performance gains due to fusion of multiple sensor modalities and evaluate both detection in single frames and performance of the full system. For the evaluation we use the dataset, which consists of daylight images, thermal images, and laser range scanner and odometry measurements collected while robot was driving along the closed path of approximately 120 meters around the office building. For the sake of single-frame evaluation, we have annotated all people appearing in the daylight images, which are larger than 40 pixels

in height. The resulting dataset contains 1480 daylight images with 300 annotated victims corresponding to 10 distinct subjects. Due to difficult illumination conditions, motion blur and large variability in viewpoints, visual people detection in such data is very challenging. At the same time detection of people in thermal images is complicated by presence of multiple background heat sources, such as heating and illumination equipment, computers and other office devices. In order to evaluate the robustness of our system to partial occlusions, we have collected an additional dataset of 28 images with 115 annotated people, 69 of which are partially occluded. We denote these datasets as “Hector Data 1” and “Hector Data 2” respectively <sup>1</sup>. In order to demonstrate the generality of our method we do not adapt visual people detector to our scenario, although this would likely lead to improved performance, and train our visual detector on the INRIA pedestrian dataset [10], where we have re-annotated the upper bodies of people. For the detection experiments we report the average precision (AP), which measures the area under the precision-recall curve. This is a common comparison measure, for which a perfect detector would achieve 100% AP. In the following we first present the results of single-frame detection of people and hazmat signs, and then evaluate performance of the full system. For the single frame detection the confidence of object hypothesis is computed according to the Eq. 11, while for the full system the confidence is based on the measurements over multiple frames.

**Single frame people detection.** Fig. 3(a) and Fig. 3(b) show the results of single-frame people detection of our system on the “Hector Data 1” and “Hector Data 2” datasets in form of recall/precision curves.

On the “Hector Data 1” dataset, detector based on visual information achieves 36.3% AP, integration of visual and laser range measurements results in 42.9% AP, and integration of visual and thermal measurements results in 43.0% AP. Integration of all three sensors leads to the best performance of 45.0% AP. The missing detections on this dataset mainly correspond to either very small or very blurry instances.

Similar trends can be observed on the “Hector Data 2”, where images contain less motion blur, but significant number of people is partially occluded. On this dataset we obtain 36.5% AP using visual detector alone, which improves to 39.1% AP by integration visual and thermal detectors. When evaluating only on the partially occluded people we obtain 27.2% AP with visual detector, and 31.1% AP with combination of visual and thermal detectors. These results show that, despite some drop in performance, our system is still producing meaningful detection results even in the case when people are partially occluded. The integration of thermal sensor measurements results in consistent improvement of performance of around 4% AP.

**Hazmat sign detection and classification.** Fig. 3(c) shows precision-recall curve quantifying single-frame hazmat sign detection performance. On this type of objects we obtain 60.1% AP. Due to smaller intra-class variability the results for hazmat signs are somewhat better than results for people detection. The missing detections are often due to motion blur and hazmat signs at extremely small scales.

We further investigate the performance of our system on hazmat sign classification task, in which the goal is to distinguish between one of the nine hazmat sign

<sup>1</sup> Both datasets are available at <http://www.gkmm.tu-darmstadt.de/rescue>

classes depicted in Fig. 3(f). For that purpose, we take the detection windows at maximum recall and assign them to one of the given classes or background.

For the classification we follow the procedure based on color histograms, described in Sec. 3. We evaluate two approaches to histogram computation, one in which color histogram is calculated on the entire detection window, and another in which detection window is subdivided into four sub-regions and separate histogram is computed for each of them. In the latter approach the final descriptor is formed by concatenating histograms of each sub-region. We obtain the recognition rate of 37.5% using histograms based on the entire window, and 58.3% using sub-region based histograms. The improvement is mainly due to better discrimination between hazmat classes with globally similar color distribution, e.g. white/red hazmat signs “Combustible” and “Flamable Solid” shown on Fig. 3(f). Region-based histograms provide better representation of the image in such difficult cases, since they are also capable of capturing the spatial distribution of colors within the detection window.

**Full system performance.** Finally, we evaluate the capability of our full system to correctly detect and localize people in the environment map. The predicted location and detection confidence of each person hypothesis is inferred by temporal integration of sensor measurements according to the filtering procedure described in Sec. 4. In contrast to single frame evaluation, the detection performance is reported for the whole series of measurements contained in the dataset. The victim is considered to be localized correctly if its predicted location on the map is within 1 meter radius the ground truth annotation, obtained by manual labeling. Multiple hits on the same ground truth annotation are only counted once, where each subsequent hit is considered a false positive.

As can be seen in Figs. 3(d) and 3(e) merging complementary information of heterogeneous devices (vis+laser+therm) outperforms all other settings by a large margin. It achieves 78.7%AP (explicit sensor fusion) and 87.2% AP (implicit sensor fusion) outperforming vis+laser by 2.9% AP and 18% AP respectively. When not using the laser, our framework suffers from placing the victims too far from ground truth annotations. vis+therm achieves 29.6% AP for explicit fusion and 49.7% AP for implicit fusion. The baseline of using only visual information achieves 29.2% AP. The implicit integration scheme achieves a higher precision for vis+thermal+laser and vis+thermal than explicit integration while the latter fusing scheme yields higher levels of recall. Note that in contrast to single-frame evaluation where recall levels are below 60%, the complete system has recall of 90% for implicit and 100% for explicit sensor fusion schemes. This is an important result for search and rescue applications, in which the ultimate goal is to find all of the victims.

## 6 Conclusion

This paper addresses sensor fusion of heterogeneous sensors with a generic semantic world model. Our framework is able to leverage complementary information for increased reliability in complex USAR scenarios. Geometric maps are enriched with semantic interpretation of scenes by detecting victims and possibly hazardous areas. The importance of sensor fusion and the expressiveness of our model are experimentally evaluated on a complex real world dataset. In future work we will address distributed sensor fusion by using multiple robots.

**Acknowledgments.** This work was supported by the DFG GRK 1362. The authors are thankful to C. Wojek for providing the GPU HOG implementation.

## References

1. Asada, M., Shirai, Y.: Building a world model for a mobile robot using dynamic semantic constraints. In: IJCAI89. (1989) 1629–1634
2. Burgard, W., Hebert, M.: World modeling. In Siciliano, B., Khatib, O., eds.: Springer Handbook of Robotics. Springer (2008) 853–869
3. Kumar, S., Guivant, J., Durrant-Whyte, H.: Informative representations of unstructured environments. In: ICRA. (2004)
4. Tadokoro, S., et al.: Robocup rescue project. *Advanced Robotics'00* **14**(5) 423–425
5. Kleiner, A., Kümmerle, R.: Genetic MRF model optimization for real-time victim detection in search and rescue. In: IROS. (2007)
6. Rottmann, A., Mozos, O.M., Stachniss, C., Burgard, W.: Semantic place classification of indoor environments with mobile robots using boosting. In: AAAI. (2005)
7. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR. (2009)
8. P. Felzenszwalb and D. McAllester and D. Ramanan: A Discriminatively Trained, Multiscale, Deformable Part Model. In: CVPR. (2008)
9. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV. (2009)
10. Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR. (2005)
11. Dalal, N. and Triggs, B. and Schmid, C.: Human detection using oriented histograms of flow and appearance. In: ECCV. (2006)
12. Ferrari, V., Marin, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR. (2009)
13. Schnitzspan, P., Fritz, M., Schiele, B.: Hierarchical Support Vector Random Fields: Joint Training to Combine Local and Global Features. In: ECCV. (2008)
14. Wojek, C., Dorkó, G., Schulz, A., Schiele, B.: Sliding-Windows for Rapid Object Class Localization: A Parallel Technique. In: DAGM-Symposium. (2008) 71–81
15. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: CVPR. (2009)
16. Jüngling, K., Arens, M.: Feature based person detection beyond the visible spectrum. In: CVPR Recognition Workshops. (2009)
17. Davis, J., Sharma, V.: Robust detection of people in thermal imagery. In: ICPR'04
18. Pham, Q.C., Gond, L., Begard, J., Allezard, N., Sayd, P.: Real-time posture analysis in a crowd using thermal imaging. In: CVPR. (2007)
19. Markov, S., Birk, A.: Detecting humans in 2d thermal images by generating 3d models. *KI 2007: Advances in Artificial Intelligence* **4667** (2007) 293–307
20. Fod, A., Howard, A., Mataric, M.J.: Laser-based people tracking. In: ICRA. (2002)
21. Arras, K., Grzonka, S., Luber, M., Burgard, W.: Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. In: ICRA. (2008)
22. Carballo, A., Ohya, A., Yuta, S.: Multiple people detection from a mobile robot using double layered laser range finders. In: ICRA Workshop. (2009)
23. Zivkovic, Z., Kröse, B.: Part based people detection using 2d range data and images. In: ICRA. (2007)
24. Gate, G., Breheret, A., Nashashibi, F.: Centralized fusion for fast people detection in dense environment. In: ICRA. (2009)
25. Schiele, B., Crowley, J.: A comparison of position estimation techniques using occupancy grids. In: ICRA. (1994)
26. Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics. MIT Press, Cambr., 2005
27. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers. (1999)