Multiclass Multimodal Detection and Tracking in Urban Environments

Luciano Spinello, Rudolph Triebel, and Roland Siegwart

Abstract This paper presents a novel approach to detect and track pedestrians and cars based on the combined information retrieved from a camera and a laser range scanner. Laser data points are classified using boosted Conditional Random Fields (CRF), while the image based detector uses an extension of the Implicit Shape Model (ISM), which learns a codebook of local descriptors from a set of hand-labeled images and uses them to vote for centers of detected objects. Our extensions to ISM include the learning of object sub-parts and template masks to obtain more distinctive votes for the particular object classes. The detections from both sensors are then fused and the objects are tracked using an Extended Kalman Filter with multiple motion models. Experiments conducted in real-world urban scenarios demonstrate the usefulness of our approach.

1 Introduction

One research area that has turned more and more into the focus of interest during the last years is the development of driver assistant systems and (semi-)autonomous cars. In particular, such systems are designed for operation in highly unstructured and dynamic environments. Especially in city centers, where many different kinds of transportation systems are encountered (walking, cycling, driving, etc.), the requirements for an autonomous system are very high. One key prerequisite for such systems is a reliable detection and distinction of dynamic objects, as well as an accurate estimation of their motion direction and speed. In this paper, we address this problem focusing on the detection and tracking of pedestrians and cars. Our system is a robotic car equipped with cameras and a 2D laser range scanner. As we will show, the use of different sensor modalities helps to improve the detection results.

Autonomous Systems Lab, ETH Zurich, Switzerland,
e-mail: {luciano.spinello, rudolph.triebel}@mavt.ethz.ch, rsiegwart@ethz.ch
This work was funded within the EU Projects BACS-FP6-IST-027140 and EUROPA-FP7-231888

The system we present here employs a variety of different methods from machine learning and computer vision, which have been shown to provide good detection rates. We extend these methods obtaining substantial improvements and combine them into a complete system of detection, sensor fusion and object tracking. We use supervised-learning techniques for both kinds of sensor modalities, which extract relevant information from large hand-labeled training data sets. In particular, the major contributions of this work are:

- Several extensions to the vision based object detector by Leibe *et al.* [13] using a feature based voting scheme denoted as Implicit Shape Models (ISM). Our major improvements to ISM are the subdivision of objects into sub-parts to obtain a more differentiated voting, the use of *template masks* to discard unlikely votes, and the definition of *superfeatures* that exhibit a higher evidence of an object's occurrence and are more likely to be found.
- The application and combination of boosted Conditional Random Fields (CRF) for classifying laser scans with the ISM based detector using vision. We use an Extended Kalman Filter (EKF) with multiple motion models to fuse the sensor information and to track the objects in the scene.

This paper is organized as follows. The next section describes work that is related to ours. Sec. 3 gives a brief overview of our overall object detection and tracking system. In Sec. 4, we introduce the implicit shape model (ISM) and present our extensions. Sec. 5 describes our classification method of 2D laser range scans based on boosted Conditional Random Fields. Then, in Sec. 6 we explain our EKF-based object tracker. Finally, we present experiments in Sec. 7 and conclude the paper.

2 Related Work

Several approaches can be found in the literature to identify a person in 2D laser data including analysis of local minima [19, 23], geometric rules [24], using maximum-likelihood estimation to detect dynamic objects [10], using AdaBoost on a set of geometrical features extracted from segments [1], or from Delaunay neighborhoods [20]. Most similar to our work is that of Douillard *et al.* [5] who use Conditional Random Fields to classify objects from a collection of laser scans. In the area of vision-based people detection, there mainly exist two kinds of approaches (see [9] for a survey). One uses the analysis of a *detection window* or *templates* [8, 4], the other performs a *parts-based* detection [6, 11]. Leibe *et al.* [13] present a people detector using *Implicit Shape Models* (ISM) with excellent detection results in crowded scenes. In earlier works, we showed already extensions of this method with a better feature selection and an improved nearest neighbor search [21, 22].

Existing people detection methods based on camera *and* laser data either use hard constrained approaches or hand tuned thresholding. Zivkovic and Kröse [25] use a learned leg detector and boosted Haar features from the camera images and employ a parts-based method. However, both their approach to cluster the laser data using

Canny edge detection and the use of Haar features to detect body parts is hardly suited for outdoor scenarios due to the highly cluttered data and the larger variation of illumination. Schulz [18] uses probabilistic exemplar models learned from training data of both sensors and applies a Rao-Blackwellized particle filter (RBPF) to track a person's appearance in the data. However, in outdoor scenarios illumination changes often and occlusions are very likely, which is why contour matching is not appropriate. Also, the RBPF is computationally demanding, especially in crowded environments. Douillard *et al.* [5] also use image features to enhance the object detection but they do not consider occlusions and multiple image detection hypotheses.

3 Overview of Our Method

Our system consists of three main components: an appearance based detector that uses the information from camera images, a 2D-laser based detector providing structural information, and a tracking module that uses the combined information from both sensor modalities and provides an estimate of the motion vector for each tracked object. The laser based detection applies a Conditional Random Field (CRF) on a boosted set of geometrical and statistical features of 2D scan points. The image based detector extends the multiclass version of the Implicit Shape Model (ISM)[13]. It only operates on a region of interest obtained from projecting the laser detection into the image to constrain the position and scale of the detected objects. Then, the tracking module applies an Extended Kalman Filter (EKF) with two different motion models, fusing the information from camera and laser. In the following, we describe the particular components in detail.

4 Appearance Based Detection

Our vision-based people detector is mostly inspired by the work of Leibe *et al.* [13] on scale-invariant Implicit Shape Models (ISM). In summary, an ISM consists in a set of local region descriptors, called the *codebook*, and a set of displacements and scale factors, usually named *votes*, for each descriptor. The idea is that each descriptor can be found at different positions inside an object and at different scales. Thus, a vote points from the position of the descriptor to the center of the object as it was found in the training data. To obtain an ISM from labeled training data, all descriptors are clustered, usually using agglomerative clustering, and the votes are computed by adding the scale and the displacement of the objects' center to the descriptors in the codebook. For the detection, new descriptors are computed on a test image and matched against the descriptors in the codebook. The votes that are cast by each matched descriptor are collected in a 3D *voting space*, and a maximum density estimator is used to find the most likely position and scale of an object.

In the past, we presented already several improvements of the standard ISM approach (see [21, 22]). Here, we show some more extensions of ISM to further improve the classification results. These extensions concern both the learning and the detection phase and are described in the following.

4.1 ISM Extensions in the Learning Phase

Sub-Parts: The aim of this procedure is to enrich the information from the voters by distinguishing between different object subparts from which the vote was cast. We achieve this by learning a circular histogram of interest points from the training data set for each object class. The number of bins of this histogram is determined automatically by using k-means clustering. The final number of clusters, here denoted as q, is obtained using the Bayesian Information Criterion (BIC). Note that this subpart extraction does not guarantee a semantical subdivision of the object (i.e.: legs, arms, etc. for pedestrians) but it is interesting to see that it nevertheless resembles this automatically without manual interaction by the user (see Fig. 1, left and center).
Template Masks: In the training data, labeled objects are represented using a binary image named *segmentation mask*. This mask has the size of the object's bounding box and is 1 inside the shape of the object and 0 elsewhere. By overlaying all these masks for a given object class so that their centers coincide and then averaging over them, we obtain a *template mask* of each object class (see Fig. 1, left and center). This method is more robust against noise than, e.g., Chamfer matching [3], and does not depend on an accurate detection of the object contours. We use the template mask later to discard outlier votes cast from unlikely areas.
Superfeatures: The original ISM maintains all features from the training data in the codebook as potential voters and does not distinguish between stronger and weaker votes. This has the disadvantage that often too many votes are cast, even if an occurance of the object is not likely given the training data, and leads to many false positive detections. To overcome this, we propose to extract *superfeatures* from the training data, i.e. descriptor vectors that cast a stronger vote than standard features. We keep these superfeatures in a separate codebook to avoid clutter in the implementation. A superfeature is defined by a local density maximum in descriptor space, where only feature vectors are considered that correspond to interest points from a dense area in the image space (in x, y, and scale). This definition ensures that for superfeatures a high evidence of the occurrence of the object is combined with a high probability to encounter an interest point. We compute superfeatures by first employing mean shift estimation on all interest points found in the training data set for each class, and then clustering the feature vectors in descriptor space that correspond to the interest points from the found areas of high density. This clustering is done agglomeratively. In the end, we select the 50% of the cluster centers that correspond to the biggest clusters. The right part of Fig. 1 shows an example. Note that the superfeatures inherently reflect the skeleton of the object.

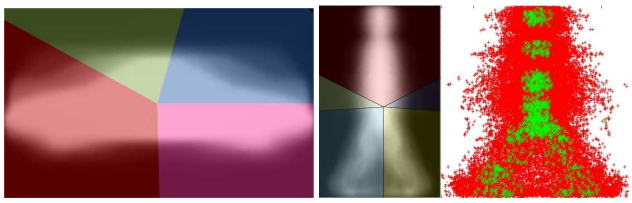

Fig. 1 Left and Center: Sub-parts, depicted in colored slices, and template masks, in white. They are computed from the training set. Note that even though the subparts are computed unsupervised, they exhibit some semantic interpretation. **Right**: Superfeatures are stable features in image and descriptor space. This figure shows Shape Context descriptors at Hessian interest points (in red) for the class 'pedestrian'. The position of the superfeatures are depicted in green.

4.2 ISM Extensions in the Inference Phase

Sub-Parts and Template Masks: After collecting all the votes for a given set of extracted input features from a test image, we first discard the ones that are implausible by placing the template mask at the potential object centers and removing the votes that are cast from outside the mask. For the remaining ones we find the maximum density point $\mathbf{m}$ using mean shift and insert all votes for $\mathbf{m}$ into a circular histogram with q bins: one per sub-part of the object. We denote each such histogram as a *hypothesis* $\mathbf{h} = (h_1, \ldots, h_q)$ of an object's position. The *strength* σ of a hypothesis is defined as the sum of all bins, i.e. the number of all voters for the object center. To find the best hypothesis we define a partial order $\prec$ based on a function Δ_h:

$$\mathbf{h}_i \prec \mathbf{h}_j \Leftrightarrow \Delta_h(\mathbf{h}_i, \mathbf{h}_j) < 0 \quad \text{where} \quad \Delta_h(\mathbf{h}_i, \mathbf{h}_j) := \sum_{k=1}^{q} \text{sign}(h_k^i - h_k^j). \tag{1}$$

Using this, we select the hypothesis with the highest order (in case of ambiguity we use the one with the highest strength) for each class. Then, we find the best hypothesis *across* all classes as described below, remove all its voters and recompute the ordering. This is done until a minimum hypothesis strength σ_{min} is reached. Thus, the parameter σ_{min} influences the number of false positive detections.

Superfeatures: Superfeatures and standard features vote for object centers in the same voting space, but the votes from superfeatures are weighted higher (in our case by a factor of 2). Thus, the score of a hypothesis is higher if the fraction of superfeatures voting for it is higher. In some cases where an object's shape visibility is low only superfeatures might be used to obtain a very fast detection.

Best Inter-Class Hypothesis: As mentioned above, we need to rate the best object hypotheses from all classes. To be independent on an over- or under-representation of a class in the codebooks, we do this by comparing the relative areas covered by the voters from all class hypotheses. More precisely, we define a square area γ around each voter that depends on the relative scale of the descriptor, i.e. the ratio of the test descriptor's scale and that of the found descriptor in the codebook. The fraction of the area covered by all voters of a hypothesis and the total area of the

object (computed from the template mask) is then used to quantify the hypothesis. Care has to be taken in the case of overlapping class hypotheses. Here, we compute the set intersection of the interest points in the overlapping area and assign their corresponding γ values alternately to one and the other hypothesis.

5 Structure Based Detection

For the detection of objects in 2D laser range scans, several approaches have been presented in the past (see for example [1, 16]). Most of them have the disadvantage that they disregard the conditional dependence between data points in a close neighborhood. In particular, they can not model the fact that the label l_i of a given scan point $\mathbf{z}_i$ is more likely to be l_j if we know that l_j is the label of $\mathbf{z}_j$ and $\mathbf{z}_j$ and $\mathbf{z}_i$ are neighbors. One way to model this conditional independence is to use Conditional Random Fields (CRFs) [12], as shown by Douillard *et al.* [5]. CRFs represent the conditional probability $p(\mathbf{y} \mid \mathbf{z})$ using an undirected cyclic graph, in which each node is associated with a hidden random variable l_i and an observation $\mathbf{z}_i$. In our case, the l_i is a discrete label that ranges over 3 different classes (pedestrian, car and background) and the observations $\mathbf{z}_i$ are 2D points in the laser scan. At this point we omit the mathematical details about CRFs and refer to the literature (e.g. [5, 17]). We only note that for training the CRF we use the L-BFGS gradient descent method [14] and for the inference we use max-product loopy belief propagation.

We use a set of statistical and geometrical features $\mathbf{f}_n$ for the nodes of the CRF, e.g. height, width, circularity, standard deviation, kurtosis, etc. (for a full list see [20]). We compute these features in a local neighborhood around each point, which we determine by jump distance clustering. However, we don't use this features directly in the CRF, because, as stated in [17] and also from our own observation, the CRF is not able to handle non-linear relations between the observations and the labels. Instead, we apply AdaBoost [7] to the node features and use the outcome as features for the CRF. For our particular classification problem with multiple classes, we train one binary AdaBoost classifier for each class against the others. As a result, we obtain for each class k a set of M weak classifiers u_i (decision stumps) and corresponding weight coefficients α_i so that the sum

$$g_k(\mathbf{z}) := \sum_{i=1}^{M} \alpha_i u_i(\mathbf{f}(\mathbf{z})) \tag{2}$$

is positive for observations assigned with the class label k and negative otherwise. We apply the inverse logit function $a(x) = (1+e^{-x})^{-1}$ to g_k to obtain a classification likelihood. Thus, the node features for a scan point $\mathbf{z}_i$ and a label l_i are computed as $\mathbf{f}_n(\mathbf{z}_i, l_i) = a(g_{l_i}(\mathbf{z}_i))$. For the edge features $\mathbf{f}_e$ we compute two values, namely the Euclidean distance d between the points $\mathbf{z}_i$ and $\mathbf{z}_j$ and a value g_{ij} defined as

$$g_{ij}(\mathbf{z}_i, \mathbf{z}_j) = \text{sign}(g_i(\mathbf{z}_i) g_j(\mathbf{z}_j))(|g_i(\mathbf{z}_i)| + |g_j(\mathbf{z}_j)|). \tag{3}$$

This feature has a high value if both $\mathbf{z}_i$ and $\mathbf{z}_j$ are equally classified (its sign is positive) and low otherwise. Its absolute value is the sum of distances from the decision boundary of AdaBoost where $g(\mathbf{z}) = 0$. Thus, we define the edge features as

$$\mathbf{f}_e(\mathbf{z}_i, \mathbf{z}_j, l_i, l_j) = \begin{cases} \begin{pmatrix} a(d(\mathbf{z}_i, \mathbf{z}_j)) & a(g_{i,j}(\mathbf{z}_i, \mathbf{z}_j)) \end{pmatrix}^T & \text{if } l_i = l_j \\ (0 \quad 0)^T & \text{otherwise.} \end{cases} \tag{4}$$

The intuition behind Eq. (4) is that edges that connect points with equal labels have a non-zero feature value and thus yield a higher potential.

6 Object Tracking and Sensor Fusion

To fuse the information from camera and laser and for object tracking we use an Extended Kalman Filter (EKF) as presented in [21]. In our implementation, we use two different motion models – Brownian motion and linear velocity – in order to cope with pedestrian and car movements. The data association is performed in the camera frame: we project the detected objects from the laser scan into the camera image. Assuming a fixed minimal object height, we obtain a rectangular search region, in which we consider all hypotheses from the vision based detector for the particular object class. Using a previously calibrated distance r_0 of an object at scale 1.0 (using the normalized training height), we can estimate the distance r_{est} of a detected object in the camera image by multiplying r_0 with the scale of the object. Then, r_{est} is compared to the measured distance r_{meas} from the laser and both detections are assigned to each other if $|r_{meas} - r_{est}|$ is smaller than a threshold τ_d (in our case $2m$).

We track cluster centers of gravity in the 2D laser frame using two system states:

$$\mathbf{x}_{m1} = \langle (x^{cog}, y^{cog}), (v_x^{cog}, v_y^{cog}), (c_1, \ldots, c_n) \rangle \text{ and } \mathbf{x}_{m2} = \langle (x^{cog}, y^{cog}), (c_1, \ldots, c_n) \rangle,$$

one for each motion model. Here, (v_x^{cog}, v_y^{cog}) is the velocity of the cluster centroid (x_x^{cog}, y_y^{cog}) and $c_1, \ldots, c_n$ are the probabilities of all n classes. We use a static state model where the observation vector $\mathbf{w}$ consists of the position of the cluster and the class probabilities for each sensor modality:

$$\mathbf{w} = \langle \hat{x}^{cog}, \hat{y}^{cog}, (c_1, \ldots, c_n)^1, \ldots, (c_1, \ldots, c_n)^s \rangle. \tag{5}$$

Here, $(\hat{x}^{cog}, \hat{y}^{cog})$ is a new observation of a cluster center and s denotes the number of sensors. The matrix H models the mapping from states to the predicted observation and is defined as $H = (P^T S_1^T \ldots S_s^T)^T$, where P maps to pose observations and the S_i map to class probabilities per sensor. For example, for one laser, one camera and constant velocity we have

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad S_1 = S_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{6}$$

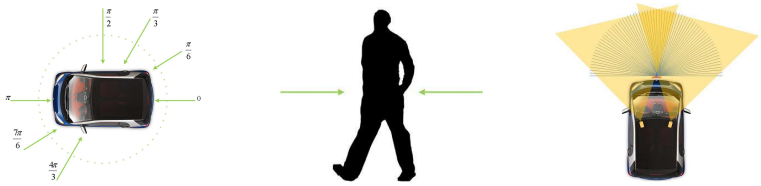

Fig. 2 Left: For car classification, we use codebooks from 7 different views. For training, mirrored images are included for each view to obtain a wider coverage. **Center:** For pedestrians we use 2 codebooks of side views with mirroring. Lateral views have sufficient information to generalize frontal/back views. **Right:** Setup used for the city data set. Only a small overlap of the cameras' field of view is used to cover a larger part of the laser scans. No stereo vision is used in this work.

7 Experimental Results

To acquire the data, we used a car equipped with two CCD cameras and a 2D laser range finder mounted in front (see Fig. 2, right). The 3D transform between the laser and the camera coordinate frame was calibrated beforehand. We acquired training data sets for both sensor modalities. For the camera, we collected images of pedestrians and cars that we labeled by hand. The pedestrian data set consists of 400 images of persons with a height of 200 pixels in different poses and with different clothing and accessories such as backpacks and hand bags in a typical urban environment. The class 'car' was learned from 7 different viewpoints as in [13] (see also Fig. 2, left). Each car data set consists of 100 pictures from urban scenes with occlusions. Car codebooks are learned using Shape Context (SC) descriptors [2] at Hessian-Laplace interest points [15]. The pedestrian codebook uses lateral views and SC descriptors at Hessian-Laplace and Harris-Laplace interest points for more robustness. Experience shows [13] that lateral views of pedestrians also generalize well to front/back views. Our laser training data consists of 800 annotated scans with pedestrians, cars and background. There is no distinction of car views in the laser data as the variation in shape is low. The range data consists in 4 layers where each has an angular resolution of 0.25° and a maximum range of $15m$.

To quantify the performance of our detector we acquired two datasets containing cars and pedestrians. The results of our detection algorithm are shown in Fig. 3. Our vision based detecion named ISMe2.0 is compared to the standard ISM, our previous extension ISMe1.0, and for the pedestrian class, with AdaBoost trained on Haar features (ABH). For the class 'car', we averaged the results over all different views. We can see that our method yields the best results with an Equal Error Rate (EER) of 72.3% for pedestrians and 74% for cars. The improvements are mainly due to a decreased rate of false positive detections. The results of our laser based detection are shown in the middle column of Fig. 3. We can see that our approach using boosted CRFs performs better than standard AdaBoost. The right column of Fig. 3 depicts the results for the combined detection using laser and vision. These graphs clearly show that using both sensors the number of false positive detections decreases and the hit rate increases. Some qualitative results are shown in Fig. 4 where a passing car and a crossing pedestrian are correctly detected and tracked.

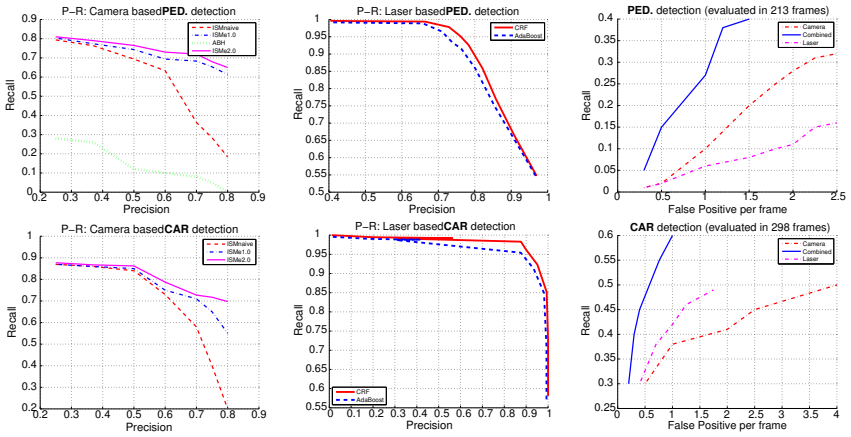

Fig. 3 Quantitative evaluation. **Upper row:** pedestrian detection, **Lower row:** car detection. From left to right we show the results only using camera, only using laser, and both. As we can see, our approach outperforms the other methods for both sensor modalities. The image based detection is compared with standard ISM, our first extension of ISM (ISMe1.0) and AdaBoost with Haar features. Our CRF-based laser detector is compared with AdaBoost. We can also see that the combination of both sensors improves the detection result of both single sensors.

In addition, we evaluated our algorithm on a third, more challenging dataset acquired in the city of Zurich. It consists of 4000 images and laser scans. The equal error rates of this experiment resulted in 64.1% (laser-only), 64.1% (vision-only) and 68% (combined) for pedestrians, and in $(72.2\%, 73.5\%, 75.7\%)$ for cars. As a comparison, we evaluated the state-of-the-art pedestrian detector based on Histogram of Oriented Gradients [4] and ABH obtained an EER of 36.4 and 8.9.

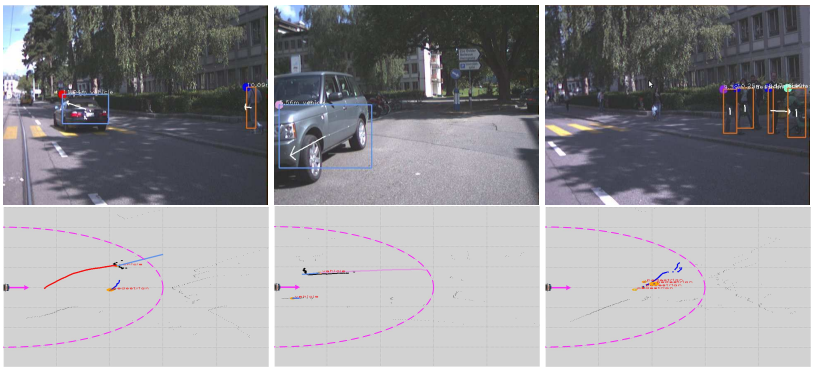

Fig. 4 Cars and pedestrian detected and tracked under occlusion, clutter and partial views. In the camera images, upper row, blue boxes indicate car detections, orange boxes pedestrian detections. The colored circle on the upper left corner of each box is the track identifier. Tracks are shown in color in the second row and plotted with respect to the robot reference frame.

8 Conclusions

We presented a method to reliably detect and track multiple object classes in outdoor scenarios using vision and 2D laser range data. We showed that the overall performance of the system is improved using a multiple-sensor system. We presented several extensions to the ISM based image detection to cope with multiple classes. We showed that laser detection based on CRFs performs better than a simpler AdaBoost classifier and presented tracking results on combined data. Finally, we showed the usefulness of our approach through experimental results on real-world data.

References

1. K. O. Arras, Ó. M. Mozos, and W. Burgard. Using boosted features for the detection of people in 2d range data. In *IEEE Int. Conf. on Rob. and Autom. (ICRA)*, 2007.
2. S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 24(4):509–522, 2002.
3. G. Borgefors. Hierarchical chamfer matching: A parametric edge matching algorithm. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 10(6):849–865, 1988.
4. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, 2005.
5. B. Douillard, D. Fox, and F. Ramos. Laser and vision based outdoor object mapping. In *Robotics: Science and Systems (RSS)*, Zurich, Switzerland, June 2008.
6. P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, pages 66–73, 2000.
7. Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
8. D. Gavrila and V. Philomin. Real-time object detection for “smart“ vehicles. In *IEEE Int. Conf. on Computer Vision (ICCV)*, 1999.
9. D. M. Gavrila. The visual analysis of human movement: A survey. *Comp. Vis. and Image Und. (CVIU)*, 73(1):82–98, 1999.
10. D. Hähnel, R. Triebel, W. Burgard, and S. Thrun. Map building with mobile robots in dynamic environments. In *IEEE Int. Conf. on Rob. and Autom. (ICRA)*, 2003.
11. S. Ioffe and D. A. Forsyth. Probabilistic methods for finding people. *Int. Journ. of Comp. Vis.*, 43(1):45–68, 2001.
12. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmentation and labeling sequence data. In *Int. Conf. on Machine Learning (ICML)*, 2001.
13. B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool. Dynamic 3d scene analysis from a moving vehicle. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, 2007.
14. D. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Programming*, 45(3, (Ser. B)), 1989.
15. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005.
16. C. Premebida, G. Monteiro, U. Nunes, and P. Peixoto. A lidar and vision-based approach for pedestrian and vehicle detection and tracking. In *ITSC*, 2007.
17. F. Ramos, D. Fox, and H. Durrant-Whyte. CRF-matching: Conditional random fields for feature-based scan matching. In *Robotics: Science and Systems (RSS)*, 2007.
18. D. Schulz. A probabilistic exemplar approach to combine laser and vision for person tracking. In *Robotics: Science and Systems (RSS)*, 2006.
19. D. Schulz, W. Burgard, D. Fox, and A. Cremers. People tracking with mobile robots using sample-based joint probabilistic data ass. filters. *Int. Journ. of Rob. Res. (IJRR)*, 22(2), 2003.

20. L. Spinello and R. Siegwart. Human detection using multimodal and multidimensional features. In *IEEE Int. Conf. on Rob. and Autom. (ICRA)*, 2008.
21. L. Spinello, R. Triebel, and R. Siegwart. Multimodal detection and tracking of pedestrians in urban environments with explicit ground plane extraction. In *IEEE Int. Conf. on Intell. Rob. and Systems (IROS)*, 2008.
22. L. Spinello, R. Triebel, and R. Siegwart. Multimodal people detection and tracking in crowded scenes. In *Proc. of the AAAI Conf. on Artificial Intelligence*, July 2008.
23. E. A. Topp and H. I. Christensen. Tracking for following and passing persons. In *IEEE Int. Conf. on Intell. Rob. and Systems (IROS)*, 2005.
24. J. Xavier, M. Pacheco, D. Castro, A. Ruano, and U. Nunes. Fast line, arc/circle and leg detection from laser scan data in a player driver. In *IEEE Int. Conf. on Rob. and Autom. (ICRA)*, 2005.
25. Z. Zivkovic and B. Kröse. Part based people detection using 2d range data and images. In *IEEE Int. Conf. on Intell. Rob. and Systems (IROS)*, San Diego, USA, November 2007.