Proceedings of the IEEE ICRA 2009
Workshop on People Detection and Tracking
Kobe, Japan, May 2009

A Trained System for Multimodal Perception in Urban Environments

Luciano Spinello, Rudolph Triebel and Roland Siegwart

Autonomous Systems Lab, ETH Zurich, Switzerland
email: {luciano.spinello, rudolph.triebel, roland.siegwart}@mavt.ethz.ch

***Abstract*— This paper presents a novel approach to detect and track multiple classes of objects based on the combined information retrieved from camera and laser rangescanner. Laser data points are classified using Conditional Random Fields (CRF) that use a set of multiclass Adaboost classified features. The image detection system is based on Implicit Shape Model (ISM) that learns an appearance codebook of local descriptors from a set of hand-labeled images of pedestrians and uses them in a voting scheme to vote for centers of detected people. We propose several extensions in the training phase in order to automatically create subparts and probabilistic shape templates, and in the testing phase in order to use these extended information to select and discriminate between hypothesis of different classes. Finally the two information are combined during tracking that is based on kalman filters with multiple motion models. Experiments conducted in real-world urban scenarios demonstrate the usefulness of our approach.**

I. Introduction

Urban environments are complex scenes where often multiple objects interact and move. In order to navigate and understand such environment a robot should be able to detect and track multiple classes of objects: most important pedestrians and cars. The ability to reliably detect these objects in real-world environments is crucial for a wide variety of applications including video surveillance and intelligent driver assistance systems. Pedestrians are particularly difficult to detect because of their high variability in appearance due to clothing, illumination and the fact that the shape characteristics depend on the view point. In addition, occlusions caused by carried items such as backpacks or briefcases, as well as clutter in crowded scenes can render this task even more complex, because they dramatically change the shape of a pedestrian. Cars are large objects that dramatically change their shape with respect to the viewpoint: for example a side view of a car is totally different from its back view. Shape symmetries can easily create false detections and shadows can drive off detection systems.

Our goal in this paper is to detect pedestrians and cars and localize them in 3D at any point in time. In particular, we want to provide a position and a motion estimate that can be used in a mobile robotic application. The real-time constraint makes this task particularly difficult and requires faster detection and tracking algorithms than the existing approaches. Our work makes a contribution into this direction. The approach we propose is multimodal in the sense that we use laser range data and images from a camera cooperatively. This has the advantage that both *geometrical structure* and *visual appearance* information are available for a more robust detection.

Managing detection of multiple classes in laser range data is a complex task due the problem of data segmentation. Often range data is grouped in consistent clusters and then classified, using heuristic rules and therefore creating a strong prior in the algorithm. In this paper, we propose an elegant solution to train and classify range data using Conditional Random Fields (CRF) through the use of a boosted set of features. Moreover each scan point will be labeled with a probability of owning to a certain class. In order to manage occlusions in complex visual scenarios a new extension of the Implicit Shape Model (ISM) for camera data classification has been developed. Finally, each detected object is tracked using a greedy data association method and multiple Extended Kalman Filters that use different motion models. This way, the filter can cope with a variety of different motion patterns for several persons simultaneously. In particular, the major contributions of this work are:

- An improved version of the image-based object detector by Leibe *et al.* [14]. It consists in several extensions to the Implicit Shape Model (ISM) in the training step, in the detection step and in the capability of coping with multiple classes. We introduce an automatic subpart extraction that is used to build an improved hypotheses selection, the concept of *superfeatures* that define a favorable feature selection that maintaining information richness. Moreover we introduce an automatically generated probability template map to ease the multiclass hypothesis selection.
- The combined use of Conditional Random Fields and camera detection to track objects in the scene.

This paper is organized as follows. The next section describes previous work that is relevant for our approach. Then, we give a brief overview of our overall object detection and tracking system. The following section presents in detail our detection method based on conditional random fields for 2D laser range data. Then, we introduce the implicit shape model (ISM) and present our extensions. Subsequently, we explain our EKF-based tracking algorithm. Finally, we present experiments and conclude the paper.

II. Previous Work

Several approaches can be found in the literature to identify a person in 2D laser data including analysis of local

minima [20], [24], geometric rules [26], or a maximum-likelihood estimation to detect dynamic objects [10], or learning AdaBoost classifiers from a set of geometrical features extracted from segments [2] or from Delaunay neighborhoods [21]. Most similar to our work is the work of [5] that makes use of a Conditional Random Field in order to label points to extract objects from a collection of laser scans.

In the area of image-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], [25], the other performs a *parts-based* detection [6], [11]. Leibe *et al.* [14] presented an image-based people detector using *Implicit Shape Models* (ISM) with excellent detection results in crowded scenes. An extension of this method that proposes a feature selection enhancement and a nearest neighbor search optimization has been already shown in [22][23].

Existing people detection methods based on camera *and* laser rangefinder data either use hard constrained approaches or hand tuned thresholding. Zivkovic and Kröse [27] use a learned leg detector and boosted Haar features extracted from the camera images to merge this information into a parts-based method. However, both the proposed approach to cluster the laser data using Canny edge detection and the extraction 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 encountered there. Therefore, we use an improved clustering method for the laser scans and SIFT features for the image-based detector. Schulz [19] uses probabilistic exemplar models learned from training data of both sensors and applies a Rao-Blackwellized particle filter (RBPF) in order to track the person's appearance in the data. However, in outdoor scenarios lighting conditions change frequently and occlusions are very likely, which is why contour matching is not appropriate. Moreover, the RBPF is computationally demanding, especially in crowded environments. The work of Douillard [5] also uses image features in order to enhance object detection but it doesn't explicitly handle occlusions and separate image detection hypotheses.

III. Overview of the method

Our system is composed 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 detection system extends the multiclass version of the Implicit Shape Model (ISM)[13] and uses Shape Context descriptors [3] computed at Harris-Laplace and Hessian interest points. It also uses the laser based detection result projected into the image to constrain the position and scale of the detected objects. Then, the tracking module applies an Extended Kalman Filter (EKF), to the combined detection results where two different motion models are implemented to account for a high variety of possible object motions. In the following, we describe the particular components in detail.

IV. Appearance Based Detection

Our image-based people detector is mostly inspired by the work of Leibe *et al.* [14] 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 of the votes is that each descriptor can be found at different positions inside an object and at different scales, and thus a vote points from the position of the descriptor to the center of the object as it was found in the training data set. To obtain an ISM from labeled training data, all descriptors are first clustered, usually using agglomerative clustering, and then 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 given 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.

A. Extensions to ISM

In the past, we presented already several improvements of the standard ISM approach (see [23], [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.

1) ISM Extensions in the Learning Phase:

a) Learning of Subparts: The aim of this procedure is to enrich the information that is obtained 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 a given object class. The number of bins of this histogram is determined automatically by using K-means clustering. The number K of clusters is obtained using the Bayesian Information Criterion (BIC). Note that this subpart extraction does not guarantee a semantical subdivision (i.e.: legs, arms in the case of pedestrians) of the object but it is interesting to see that it nevertheless resembles this automatically without manual interaction by the user (see Fig. 1, left).

b) Applying a Template Mask: The idea here is to extract a common segmentation mask from the training data for each object by averaging over all masks from the particular object instances. This mask is later used to discard outlier voters by overlaying the mask at the hypothetical center of the object. Chamfer matching has been widely used in literature [4] to compute such a mask. However, it heavily depends on a robust detection of the contour edges and is strongly affected by noise. A more robust method is to build a

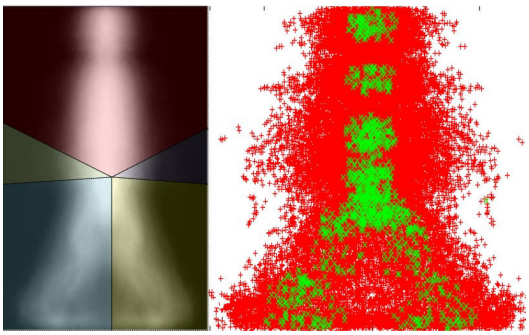

Fig. 1. **Left**: Probabilistic template and overlayed subparts are both automatically computed from the training set (in this case on the class 'pedestrian'). It is important to notice that even though the subparts are computed without a semantic subdivision, their segmentation shows legs, arms and upper body.**Right**: Superfeatures are stable features in image and descriptor space. This figure depicts Shape Context descriptors with Hessian Interest point (in red) in the case of pedestrian class. In green are depicted the selected superfeatures.

probabilistic template map from the individual segmentation masks in the training set. All the segmentation masks are collected, centered with respect to their center of gravity and averaged. Strong responses (common areas of the same objects) have high probability, whereas various details are softened in the average but still kept.

2) Learning Superfeatures: The original ISM does not perform feature selection but it maintains the complete probability distribution generated by extracted features of the training set. This has the disadvantage to potentially generate false positive due to inevitable feature mismatches. We here propose a method to drive the detection while still maintaining information richness. The idea is to find good features in the image space (namely $\langle x, y, scale \rangle$) and descriptor space ($n$-d space) that could vote for the object center with more weight to ease the hypothesis selection. The procedure can be sketched in three steps.

1) Interest points of the entire training dataset are collected.
2) Dense areas of interest points reflect a high informative content. We employ mean shift mode seeking with a uniform kernel in order to locate such areas.
3) On each convergence point descriptors are collected in pools. These pools are clustered using unsupervised clustering with average linkage in order to group closely similar features. We use the best 50% of the resulting groups (ranked by quantity) and collect them as *superfeatures*.

Noticeably, the resulting superfeatures inherently reflect the skeleton of the objects and constitute key points in the shape of the objects (see Fig. 1, right).

B. ISMe: extensions in testing phase

In this subsection we explain how we combine the richer learning information in order to obtain a better detection.

1) Using superfeatures: Superfeatures and features vote for object centers in the same voting space: the votes generated by the first are bigger than the latter. The resulting hypothesis score is enriched by their support. In visually simple scenes it is possible to apply just superfeature codebooks in order to obtain a very fast detection.

2) Using subparts and prob. template in the cost function: Each hypothesis is now defined by an angular histogram in which the bins are defined by the subparts. Moreover, the probabilistic template is used to prune feature matches that lie far outside the probabilistic shape (that is scaled according to the hypothesis). In order to determine which of the hypotheses better represents an object of a given class, we use a maximum likelihood estimation method. In particular, we solve:

$$\mathcal{H}_s = \operatorname{argmax} p(\mathcal{H}|\Theta), \tag{1}$$

where $\mathcal{H}$ represents the set of hypotheses and Θ is the feature assignment. In order to achieve the solution we consider pairwise comparisons. Given each pair of hypotheses h_a and h_b, their relative histograms $\mathcal{W}_a = \{w_1^a, ..., w_q^a\}$, and $\mathcal{W}_b$ we compute:

$$V = \sum_i^q v_i \tag{2}$$

where

$$v_i = \begin{cases} 1 & \text{if} \quad w_i^a > w_i^b \\ -1 & \text{otherwise} \end{cases} \tag{3}$$

Then a simple sign condition is used to check which of the two hypotheses is the best. If we perform this simple and fast comparison on the set $\mathcal{H}$, we obtain $h_{max} = \operatorname{argmax} p(\mathcal{H}|\Theta)$ and put it in the selected hypothesis set $\mathcal{H}_c$.

3) Discriminate between object classes: In the previous subsection we explained how we selected the best object hypothesis for each class. Here we explain how we discriminate among hypothesis of different classes. In order to not bias the multiclass detection towards a class that has more features or codebook occurrences we used a common measure to do hypothesis selection. This comes from the probabilistic template area ratio. Each assigned feature for a certain hypothesis occupies a scaled square area in the probabilistic template. The ratio of the occupied area on the total object area is the score of each class hypothesis. For each object class hypothesis a score s_i is computed taking into account the overlapping area (if present) between hypotheses of different classes:

$$s_i = r_i - \frac{\Delta o}{\#o} \tag{4}$$

where r_i is the area ratio and Δo is the overlap ratio of the areas, and $\#o$ is the number of overlaps. The best score s_i defines the current winning object hypothesis. The features involved in the voting of this hypothesis are then removed from the voting space and the selection process (subparts voting and object class selection) continues until a detection with a minimum strength d_t is available.

This two step process is necessary to handle occlusions and multiple classes in a computationally feasible time: each hypothesis competes with the rest of its class to become the best hypothesis of its class. Then it is evaluated against all

the other candidates of the other class and then, if it is the case, selected.

V. Structure Based Detection

For the detection of objects in 2D laser range scans, several approaches have been presented in the past. Most of theses approaches have the disadvantage that they disregard the conditional dependence between data points in a close neighborhood: the fact that the label y_i of a given scan point $\mathbf{z}_i$ is more likely to be y_j if we know that y_j is the label of $\mathbf{z}_i$'s neighbor $\mathbf{z}_j$ is not reflected. One way to model this conditional independence is to use Conditional Random Fields (CRFs) [12], as has been 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 y_i and an observation $\mathbf{z}_i$. In our case, the y_i is a discrete label that ranges over 2 different classes (pedestrian and car) and the observations $\mathbf{z}_i$ are 2D points in the laser scan. Assuming a maximal clique size of 2 for the graph, we can compute the conditional probability of the labels $\mathbf{y}$ given the observations $\mathbf{z}$ as:

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{1}{Z(\mathbf{z})} \prod_{i=1}^{N} \varphi(\mathbf{z}_i, y_i) \prod_{(i,j)\in\mathcal{E}} \psi(\mathbf{z}_i, \mathbf{z}_j, y_i, y_j), \quad (5)$$

where $Z(\mathbf{z}) = \sum_{\mathbf{y}'} \prod_{i=1}^{N} \varphi(\mathbf{z}_i, y_i') \prod_{(ij)\in\mathcal{E}} \psi(\mathbf{z}_i, y_i', y_j')$ is usually called the *partition function* and $\mathcal{E}$ is the set of edges in the graph. To determine the node and edge potentials φ and ψ we use the log-linear model:

$$\varphi(\mathbf{z}_i, y_i) = e^{\mathbf{w}_n \cdot \mathbf{f}_n(\mathbf{z}_i, y_i)},\ \psi(\mathbf{z}_i, \mathbf{z}_j, y_i, y_j) = e^{\mathbf{w}_e \cdot \mathbf{f}_e(\mathbf{z}_i, \mathbf{z}_j, y_i, y_j)}$$

where $\mathbf{f}_n$ and $\mathbf{f}_e$ are feature functions for the nodes and the edges in the graph, and $\mathbf{w}_n$ and $\mathbf{w}_e$ are the feature weights that are determined in the training phase. The computation of the partition function Z is intractable due to the exponential number of possible labelings $\mathbf{y}'$. Instead, we compute the *pseudo-likelihood*, which approximates $p(\mathbf{y} \mid \mathbf{z})$ and is defined by the product of all likelihoods computed on the *markov blanket* (direct neighbors) of node i.

$$pl(\mathbf{y} \mid \mathbf{z}) = \prod_{i=1}^{N} \frac{\varphi(\mathbf{z}_i, y_i) \prod_{\mathbf{z}_j \in \mathcal{N}(\mathbf{z}_i)} \psi(\mathbf{z}_j, \mathbf{z}_i, y_j, y_i)}{\sum_{\mathbf{y}'} \left(\varphi(\mathbf{z}_i, y_i') \prod_{\mathbf{z}_j \in \mathcal{N}(\mathbf{z}_i)} \psi(\mathbf{z}_j, \mathbf{z}_i, y_i', y_j') \right)}$$

Here, $\mathcal{N}(\mathbf{z}_i)$ denotes the set of direct neighbors of node i. In the training phase, we compute the weights $\mathbf{w}_n$ and $\mathbf{w}_e$ that minimize the negative log pseudo-likelihood together with a Gaussian shrinkage prior as in [18]:

$$L(\mathbf{w}) = -\log pl(\mathbf{y} \mid \mathbf{z}) + \frac{(\mathbf{w} - \hat{\mathbf{w}})^T (\mathbf{w} - \hat{\mathbf{w}})}{2\sigma^2} \quad (6)$$

For the minimization of L, we use the L-BFGS gradient descent method [15]. Once the weights are obtained, they are used in the inference phase to find the labels $\mathbf{y}$ that maximize Eq. (5). Here, we do not need to compute the partition function Z, as it is not dependent on $\mathbf{y}$. We use max-product loopy belief propagation to find the distributions of each label y_i. The final labels are then obtained as those that are most likely for each node.

A. Node and Edge Features

As node features $\mathbf{f}_n$ we use a set of statistical and geometrical features such as height, width, circularity, standard deviation, kurtosis, etc. (for a full list see [21]). We compute these features in a local neighborhood around each point, which we determine by jump distance clustering. We can then use these features as an input to the CRF classification algorithm. However as stated in [18], and also from our own observation, the CRF is not able to handle non-linear relations between the observations and the labels, which is a consequence of the log-linear model described above. To overcome this problem, we apply AdaBoost [7] to the node features and use the outcome of AdaBoost 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 a set of weak classifiers h_i (decision stumps) and corresponding weight coefficients α_i so that the sum

$$g_k(\mathbf{z}) = \sum_{i=1}^{M} \alpha_i h_i(\mathbf{f}(\mathbf{z})) \quad (7)$$

is positive for observations that are assigned with the class label k and negative otherwise. To obtain values between 0 and 1 we apply the inverse logit function $l(x) = (1 + \exp(-x))^{-1}$, which has a sigmoid shape and ranges between 0 and 1, to each value g_j. We do this for two reasons: First we obtain values that can be interpreted as likelihoods of corresponding to class k. Second, by applying the same technique also for the edge features, the resulting potentials are better comparable. The resulting node features are then computed as

$$\mathbf{f}_n(\mathbf{z}_i, y_i) = l(g_{y_i}(\mathbf{z}_i)), \quad (8)$$

i.e. the scalar component of the vector $l(\mathbf{g})$ that corresponds to the class with label y_i. For the edge features, we don't apply AdaBoost, but instead compute two values, namely the Euclidean distance d_{ij} between the points $\mathbf{z}_i$ and $\mathbf{z}_j$ and a value g_{ij} defined as

$$g_{ij}(\mathbf{z}_i, \mathbf{z}_j) = \text{sign}(g_i(\mathbf{z}_i) g_j(\mathbf{z}_j))(|g_i(\mathbf{z}_i)| + |g_j(\mathbf{z}_j)|) \quad (9)$$

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

$$\mathbf{f}_e(\mathbf{z}_i, \mathbf{z}_j, y_i, y_j) = \begin{cases} (l(d_{y_i,y_j}) \quad l(g_{y_i,y_j})^T & \text{if } y_i = y_j \\ (0 \quad 0)^T & \text{otherwise} \end{cases} \quad (10)$$

Here, we omitted the arguments $\mathbf{z}_i$ and $\mathbf{z}_j$ of the functions d_{ij} and g_{ij} for brevity. The intuition behind Eq. (10) is that

edges that connect points with equal labels have a non-zero feature value and thus yield a higher potential. The latter is sometimes referred to as the generalized Potts model (see [1], [17]).

a) Connectivity: Nowadays many laser scanners have multilayer scanning capabilities. The CRF connectivity is defined by a separate Delaunay triangulation for each layer. Between layers connectivity is assured by connecting points located in the same vertical. This assures a good layer connection for the flow of BP and lessen the arc count with respect to a full triangulation.

VI. Tracking objects for sensor fusion

In order to fuse the information coming from both sensors (camera and laser) and to simultaneously keep track of the object we use an EKF based tracking system, first introduced in [23]. Here, each object is tracked with several motion models (in this case: brownian motion and linear velocity) in order to cope with pedestrian and car movements. We perform tracking in the laser data, therefore camera detections are projected and assigned to segments in the laser data. In order to reliably track wide objects, like cars, tracking single segments are not enough. Single segments tend to be spatially very unstable due to the noise present in outdoor environments and the scatter resulting from the distance with respect to the observer. We therefore group segments with the same class label using Delaunay triangulation and a trim distance rule. The resulting cluster will have a more stable position and a probability of being a class that is the average of its members. Each Kalman filter state ($\langle x, y, (v_x, v_y)\rangle$) is augmented with N states where N is the number of classes present in the detector. Indeed, the observation vector z fed to the tracking system consists of the position of the cluster and the class label probability. The matrix H that models the observations to mapping in the Kalman Filter $x = Hz$ is defined by $H = [H_{lsr}; H_{cam}]$ in order to manage multiple inputs from different sensors.

VII. Experimental Results

A car equipped with several active and passive sensors is used to acquire the datasets. In particular, we use a monocular camera in combination with a 2D laser range finder in front of the car. An accurate camera-laser synchronization and calibration has been developed for this work.

A. Image training datasets

The scope of this paper is to detect pedestrians and cars, we therefore used a pedestrian dataset and three different datasets for cars: front view, side view, back view. The class car itself consists in multiple classes because of its different visual appearance with respect to the viewpoint. The pedestrian dataset consists of 400 images of persons with a height of 200 pixels at different positions and dressed with different clothing and accessories such as backpacks and hand bags in a typical urban environment. Each car dataset consists in a set of 100 pictures taken in several urban scenes with occlusions due to people or traffic signs.

B. Laser training datasets

The laser detector has been trained using 203 annotated laser scans containing clutter, pedestrians and cars. There is not distinction between car views in the laser detector due to a not dramatic viewpoint change in the range data. The range data is organized in 4 layers with a relative orientation of $0.8°$. Each layer has a resolution of $0.25\circ$ and maximum range of $30m$.

C. Qualitative and quantitative multiclass results

In order to determine the performance of our detector we created two datasets consisting of cars and pedestrians. The image based detection uses Shape context descriptors [3] from Hessian-Laplace and Harris-Laplace [16] interest point. The quantitative results of the performance of pedestrian based image detection are shown in the precision-recall graph of Fig. 2-left. In the graph is shown a comparison with respect to a naive ISM implementation that does not uses hypothesis selection, Adaboost based Haar detector and our previous version of the image detector (labeled as ISMe1.0). The performance increase of our approach is mainly related to the introduction of the new hypothesis selection system. The performance of image based detection for cars is shown in Fig. 2-middle where a comparison with ISM and ISMe1.0 is shown. For clarity the results are averaged between the three different views of the class car. In general we can notice from the results that pedestrian classification is harder than car classification due to shape complexity and flexibility.

In order to justify our approach for laser range data detection we evaluated CRF against Boost classifier that uses the same set of features, the resulting precision-recall graph is shown in Fig. 2-right for pedestrian and in fig. 3-left for cars. Then we evaluated the current performance of combining the information togheter. A very informative way of showing the potential of our method is shown in the two graph of Fig. 3-middle and Fig. 3-right in which we show that combining the two information increases the hit rate and decreases the false positives. We show some qualitative results extracted from the testing datasets in Fig. 4-right in which cars are correctly detected from both sensors but the pedestrian is detected only with the laser and not with the camera due to its pose configuration and its visual neighborhood. Another result is shown in Fig. 4-left in which the camera classifier detects a false positive located on vertical structures of the trolleybus and detects the person on the scooter as a pedestrian due its visual similarity. Thanks to the structure information obtained from the laser the system can discriminate the false positive. Moreover, we show qualitative tracking results in Fig. 5, Fig. 7, Fig. 6 where passing cars and a crossing pedestrian are correctly tracked using multiple sensor information.

VIII. Conclusions

In this paper we presented a method to reliably detect and track multiple classes (cars and pedestrian) in outdoor scenarios using 2D laser range data and camera images. We showed that the overall performance of the system is

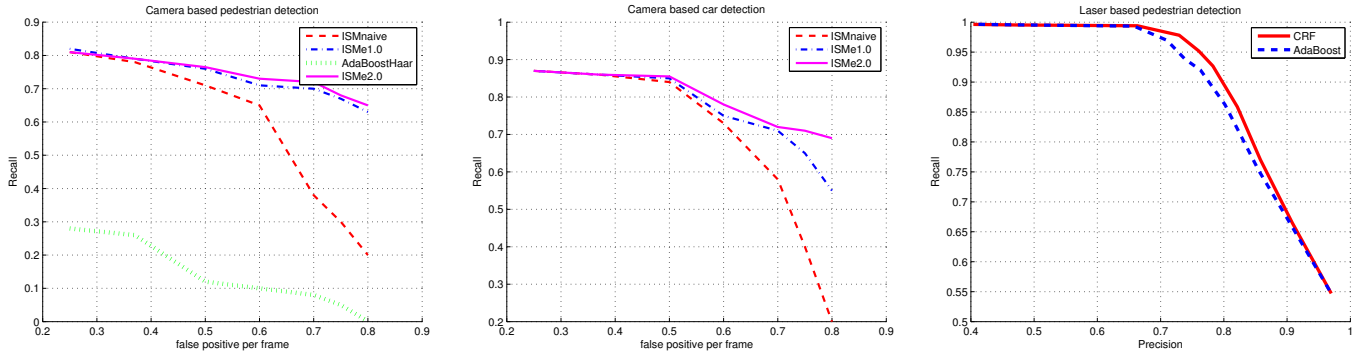

Fig. 2. **Left**:Precision-recall graph for image based pedestrian detection. Performance comparison is shown between ISM, our previous extension ISMe1.0 and Adaboost Haar based detector. **Middle**: Precision-recall graph for image based car detection. Performance comparison is shown between ISM, our previous extension ISMe1.0 and ISM **Right**: Precision-recall graph for laser range data based pedestrian detection. Performance comparison is shown between CRF and a Boost based approach

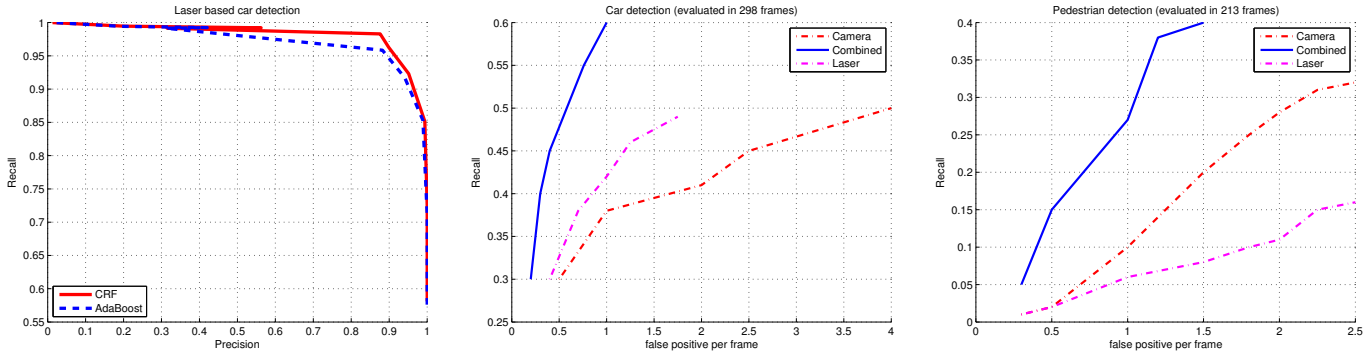

Fig. 3. **Left**: Precision-recall graph for laser range data based car detection. Performance comparison is shown between CRF and a Boost based approach. **Middle**: Recall-false positive per frame for camera-laser information fusion for car detection. A comparison of camera and laser is shown in figure. **Right** Recall-false positive per frame for camera-laser information fusion for pedestrian detection. A comparison of camera and laser is shown in figure.

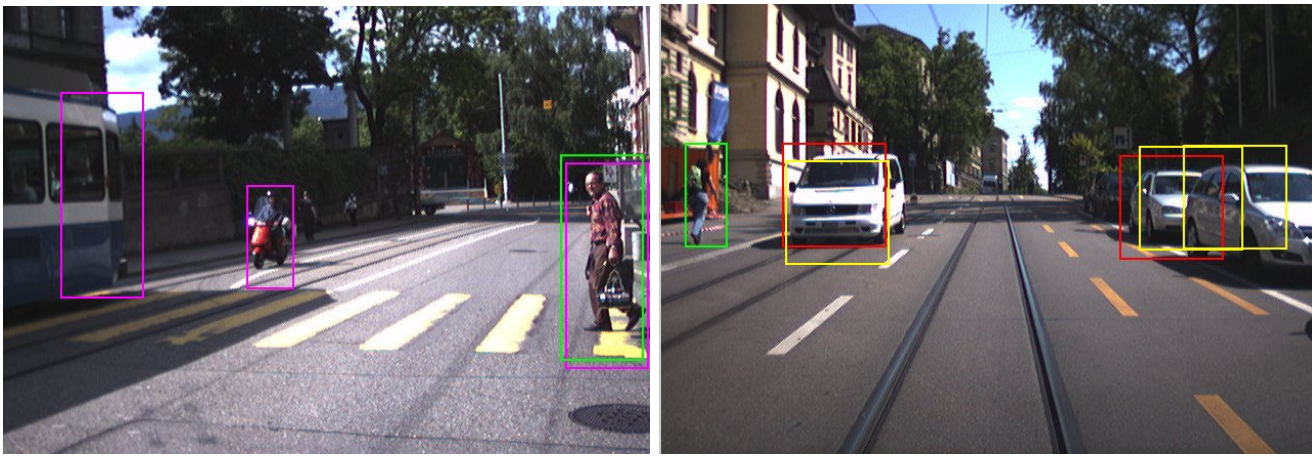

Fig. 4. Detections from multiple sensors. Green: laser based pedestrian detections; Yellow: laser based car detections; Magenta: camera based pedestrian detection; Red: camera based car detection

Fig. 5. Tracking cars in an intersection. A bounding box surrounds the tracked object with annotated distance and a colored marker that refers to the track in the laser plane.

Fig. 6. Tracking a pedestrian that crosses the road. A bounding box surrounds the tracked object with annotated distance and a colored marker that refers to the track in the laser plane. In the laser plane it is visible a false track associated with one steady detection of a cylinder concrete by the laser based detector. For clarity, the laser tracked cluster is plotted into the image (green points).

Fig. 7. Tracking cars in an intersection. A bounding box surrounds the tracked object with annotated distance and a colored marker that refers to the track in the laser plane. It is important to notice that also in case of the extreme closeup of the truck the track is still maintained

improved using a multiple sensor system. We presented several novel extensions to the ISM-based image detection in order to cope with multiple classes. We showed that a system based on CRF has better performance than a simpler Adaboost based classifier and presented tracking results on combined data. Finally, we presented experimental results on real-world data that point out the usefulness of our approach.

REFERENCES

[1] D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, and A. Ng. Discriminative learning of markov random fields for segmentation of 3d scan data. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, pages 169–176, 2005.

[2] 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. & Autom. (ICRA)*, 2007.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. volume 24, pages 509–522, 2002.

[4] G. Borgefors. Hierarchical chamfer matching: A parametric edge matching algorithm. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 10.

[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. & 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 *Proc. of the 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] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, pages 878–885, Washington, DC, USA, 2005. IEEE Computer Society.

[15] D. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Programming*, 45(3, (Ser. B)), 1989.

[16] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005.

[17] R. B. Potts. Some generalized order-disorder transformations. *Proc. Cambridge Phil Soc.*, 48, 1952.

[18] F. Ramos, D. Fox, and H. Durrant-Whyte. Crf-matching: Conditional random fields for feature-based scan matching. In *RSS*, 2007.

[19] D. Schulz. A probabilistic exemplar approach to combine laser and vision for person tracking. In *Robotics: Science and Systems (RSS)*, Philadelphia, USA, August 2006.

[20] D. Schulz, W. Burgard, D. Fox, and A. Cremers. People tracking with mobile robots using sample-based joint probabilistic data association filters. *Int. Journ. of Robotics Research (IJRR)*, 22(2):99–116, 2003.

[21] L. Spinello and R. Siegwart. Human detection using multimodal and multidimensional features. In *IEEE Int. Conf. on Rob. & Autom. (ICRA)*, 2008.

[22] 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 Sys. (IROS)*, 2008.

[23] L. Spinello, R. Triebel, and R. Siegwart. Multimodal people detection and tracking in crowded scenes. In *Proc. of The AAAI Conference on Artificial Intelligence*, July 2008.

[24] E. A. Topp and H. I. Christensen. Tracking for following and passing persons. In *IEEE Int. Conf. on Intell. Rob. and Sys. (IROS)*, 2005.

[25] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *IEEE Int. Conf. on Computer Vision (ICCV)*, page 734, Washington, DC, USA, 2003. IEEE Computer Society.

[26] 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. & Autom. (ICRA)*, pages 3930–3935, 2005.

[27] Z. Zivkovic and B. Kröse. Part based people detection using 2d range data and images. In *IEEE Int. Conf. on Intell. Rob. and Sys. (IROS)*, San Diego, USA, November 2007.