Multimodal Detection and Tracking of Pedestrians in Urban Environments with Explicit Ground Plane Extraction

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 people detection and tracking method based on a combined multimodal sensor approach that utilizes 2D and 3D laser range and camera data. Laser data points are clustered and classified with a set of geometrical features using an SVM AdaBoost method. The clusters define a region of interest in the image that is adjusted using the ground plane information extracted from the 3D laser. In this areas a novel vision based people detector based on Implicit Shape Model (ISM) is applied. Each detected person is tracked using a greedy data association technique and multiple Extended Kalman Filters that use different motion models. This way, the filter can cope with a variety of different motion patterns. The tracker is asynchronously updated by the detections from the laser and the camera data. Experiments conducted in real-world outdoor scenarios with crowds of pedestrians demonstrate the usefulness of our approach.**

I. Introduction

The ability to reliably detect people in real-world environments is crucial for a wide variety of applications including video surveillance and intelligent driver assistance systems. The detection of pedestrians is the next logical step after the development of a successful navigation and obstacle avoidance algorithm for urban environments. However, 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.

Our goal is to detect pedestrians 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 real-time 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. In this paper, we propose to exploit this information using supervised learning techniques that are based on a combination of AdaBoost with Support Vector Machines (SVMs) for the laser data and on an extension of the Implicit Shape Model (ISM) for the camera data. In the detection phase, both classifiers yield likelihoods of detecting people which are fused into an overall detection probability. The information extracted from 3D and 2D data define the positioning of the hypotheses in the image. The image detection method is constrained in region of interest generated by the 2D laser and positioned in the image using a ground plane extraction method from 3D scans. Finally, each detected person 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. The tracker is asynchronously updated by the detections from the laser and the camera data. In particular, the major contributions of this work are:

- An improved version of the image-based people detector by Leibe *et al.* [12]. It consists in three extensions to the Implicit Shape Model (ISM), resulting in a reduced computation time and an improved feature selection.
- A method to discard false positive detections by computing regions of interest in the camera images.
- The use of a 3D scanning device, which facilitates a fast and robust detection of the ground plane and thus helps to disambiguate possible detections of pedestrians.

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 people detection and tracking system. The following section presents in detail our detection method based on the 2D laser range data and explains 3D plane extraction. Then, we introduce the implicit shape model (ISM), present our extensions to the ISM and expose the region of interest generation algorithm. Subsequently, we explain our EKF-based tracking algorithm focusing particularly on the multiple motion models we use. 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 [17], [20], geometric rules [23], or a maximum-likelihood estimation to detect dynamic objects [10]. Most similar to our work is the approach of Arras *et al.* [2] which clusters the laser data and learns an AdaBoost classifier from a set of geometrical features extracted from the clusters. Recently, we extended this approach *et al.* [18] in such a way that multi-dimensional features are used and that they are learned using a cascade of Support Vector Machines (SVM)

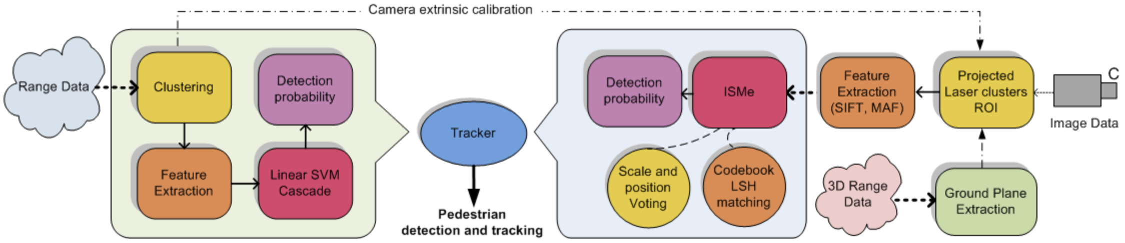

Fig. 1. Overview of the individual steps of our system. See text for details.

instead of the AdaBoost decision stumps. In the area of image-based people detection, there mainly exist two kinds of approaches (see [8] for a survey). One uses the analysis of a *detection window* or *templates* [7], [22], the other performs a *parts-based* detection [5], [11]. Leibe *et al.* [12] presented an image-based people detector using *Implicit Shape Models* (ISM) with excellent detection results in crowded scenes.

Existing people detection methods based on camera *and* laser rangefinder data either use hard constrained approaches or hand tuned thresholding. Zivkovic and Kröse [24] 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 [16] 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.

III. Overview of the Method

Our system is divided into three phases: training, detection and tracking (see Figure 1). In the training phase, the system learns a structure-based classifier from a hand-labeled set of 2D laser range scans, and an appearance-based classifier from a set of labeled camera images. The first one uses a boosted cascade of linear SVMs, while the latter computes an implicit shape model (ISM), in which a collected set of image descriptors from the training set vote for the occurrence of a person in the test set. In the detection phase, the laser-based classifier is applied to the clusters found in a new range scan and a probability is computed for each cluster to correspond to a person. The clusters are then projected into the camera image to define a region of interest and positioned using the information of the ground plane extracted from the online retrieved 3D point cloud. Thus an appearance-based classifier extracts local image descriptors and uses them to obtain a set of hypotheses of detected persons. Here, we apply a new technique to discard false positive detections. Finally in the tracking phase, the information from both classifiers is used to track the position of the people in the scan data. The tracker is updated whenever a new image or a laser measurement is received and processed. It applies several motion models per track to account for the high variety of possible motions a person can perform. For the scope of this paper, we omit the details of our tracking algorithm and refer instead to[19] for an extensive explanation. In the following, we describe the particular steps of our system in detail.

IV. Structure Information: Laser Data Analysis

Our robotic system features a 2D and a 3D laser range scanner. The dense and frequent 2D range data is used to estimate possible locations of a person's legs, and the 3D point clouds are used to extract the ground plane to aid the appearance-based person detector (see section V).

A. Clustering and Classification of 2D range data

A graph based reasoning on the classic *jump distances* segmentation has been proposed in [18] in order to address the problem of clustering range data in outdoor scenario. Experimental results showed that this reduces the cluster quantity of $25\% - 60\%$, significantly reducing overclustering but mantaining clusters information.

We use an improved version of Adaboost [6] based on a cascade of support vector machines (SVMs)[18] to classify the clustered laser data into the classes "person" and "no person". The main reason for this is to obtain a small number of classifiers in each stage and to guarantee an optimal separation of the two classes. We denote the detection of a person using a binary random variable π that is true whenever a person is detected. Each of the L cascaded SVM-classifiers h_i yields either 1 or 0 for a given input feature vector $\mathbf{f}$. The overall detection probability can then be formulated as

$$p(\pi \mid \mathbf{f}) = \sum_{i=1}^{L} w_i h_i(\mathbf{f}) \tag{1}$$

In the learning phase, the weights w_i and the hyperplanes are computed for each SVM classifier h_i. The laser-based people detector then computes (1) for each feature vector $\mathbf{f}$ in the test data set.

B. Ground Plane Extraction from 3D Scans

As mentioned, a point cloud $\mathcal{P}$ obtained with our 3D rotating scanner device reflects the full 360° environment of the vehicle. The idea is to use this information to extract the

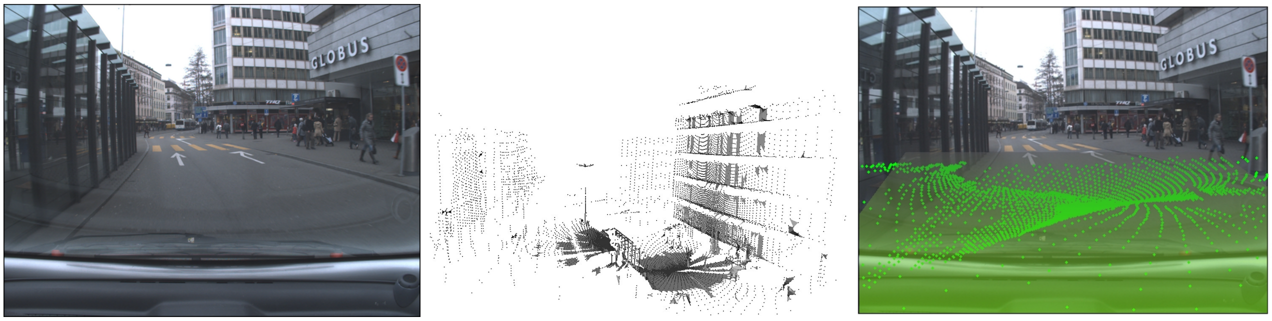

Fig. 2. 3D ground plane extraction. **Left:** Camera image as seen from the inside of the vehicle. **Middle:** Triangulated 3D point cloud of the same scene (seen from above). **Right:** Camera image with the points of the extracted ground plane overlayed.

position of the ground plane in the local environment of the vehicle to be able to further disambiguate detected persons from the camera images and to reduce false positives. In the literature, there exist many different approaches to detect planes in 3D range data [13], [21], [9]. For the application described here we want to detect and track persons if as fast as possible. Therefore, we decided to use a simple but time efficient region growing technique to detect the ground plane. The criterion for a scan point to belong to the ground plane is that its corresponding normal vector deviates only slightly (in our implementation by maximal $25°$) from the upright vector $(0, 0, 1)^T$ and that it is not farther away from its closest neighbor than a given threshold (we use $1\ m$). The region growing is initiated always at the same fixed point right in front of the vehicle at the ground level. To efficiently compute the normal vectors, we exploit the fact that the point clouds are structured in *slices* – each scan line of the vertically mounted rotating laser scanner accounts for one slice. This facilitates a fast and simple mesh triangulation performed by connecting two consecutive points from one slice with one point of the consecutive slice. From this triangulation the normal vectors are easily computed from the normalized cross product of difference vectors. An example result of the ground plane extraction is shown in Figure 2. To clarify: rectangular bounding boxes are created in the image where laser clusters are found then the extracted ground plane is used to place those region of intereset (ROI) at the correct height in the image. The resulting ROI placement helps the image detector in creating valid detection hypotheses.

V. Appearance Information: Image Data Analysis

Our image-based people detector is mostly inspired by the work of [12] on scale-invariant Implicit Shape Models (ISM). An ISM is a generative model for object detection. In this paper we extend this approach, but before we briefly explain the steps for learning an object model in the original ISM framework.

An Implicit Shape model consists of a *codebook* $\mathcal{I}$ and a set of votes $\mathcal{V}$. The K elements of $\mathcal{I}$ are local region descriptors $\mathbf{d}_1^C, \ldots, \mathbf{d}_K^C$ and $\mathcal{V}$ contains for each $\mathbf{d}_i^C$ a set of D_i local displacements $\{(\Delta x_{i,j}, \Delta y_{i,j})\}$ and scale factors $\{s_{i,j}\}$ with $j = 1, \ldots, D_i$. The interpretation of the votes is that each descriptor $\mathbf{d}_i^C$ can be found at different positions inside an object and at different scales. To account for this, each local displacement points from $\mathbf{d}_i^C$ to the center of the object as it was found in the labeled training data set. To obtain an ISM from a given training data set, two steps are performed:

1) **Clustering** All region descriptors are collected from the training data. The descriptors are then clustered using agglomerative clustering with average linkage. In the codebook, only the cluster centers are stored.
2) **Computing Votes** In a second run over the training data, the codebook descriptors $\mathbf{d}_i^C$ are matched to the descriptors $\mathbf{d}_j^I$ found in the images, and the scale and center displacement corresponding to $\mathbf{d}_j^I$ is added as a vote for $\mathbf{d}_i^C$.

In the detection phase, we again compute interest points $\mathbf{x}_j^I$ and corresponding region descriptors $\mathbf{d}_j^I$ at various scales on a given test image I. The descriptors are matched to the codebook and a matching probability $p(\mathbf{d}_i^C \mid \mathbf{d}_j^I)$ is obtained for each codebook entry. With the sample-based representation, we can detect a person at location $\bar{\mathbf{x}}$ by a maxima search using variable bandwidth mean shift balloon density estimation [4] in the 3D voting space.

A. First Extension to ISM: Strength of Hypotheses

In the definition of the ISM there is no assumption made on the particular shape of the objects to be detected. This has the big advantage that the learned objects are detected although they might be occluded by other objects in the scene. However, the drawback is that usually there is a large number of false positive detections in the image background. [12] address this problem using a minimum description length (MDL) optimization based on pixel probability values. However, this approach is rather time demanding and not suited for real-time applications. Therefore, we suggest a different approach.

First we evaluate the quality of a hypothesis of a detected object center $\mathbf{x}$ with respect to two aspects: the overall *strength* of all votes and the way in which the voters are *distributed*. Assume that ISM yields an estimate of a person at position $\mathbf{x}$. We can estimate the spatial distribution of voters $\mathbf{x}_j^I$ that vote for $\mathbf{x}$ using a 1D circular histogram that ranges from 0 to 2π. When computing the weight of the vote we also compute the angle α

$$\alpha(\mathbf{x}_j^I, \mathbf{x}) = \arctan2(y_j^I - y, x_j^I - x) \tag{2}$$

and store the voting weight in the bin that corresponds to α. This way we obtain a histogram $\xi(\mathbf{x})$ with, say, B bins for each center hypothesis $\mathbf{x}$. Now we can define an ordering on the hypotheses based on the histogram difference:

$$d(\mathbf{x}_1, \mathbf{x}_2) := \sum_{b=1}^{B} \xi_b(\mathbf{x}_1) - \xi_b(\mathbf{x}_2), \tag{3}$$

where $\xi_b(\mathbf{x}_1)$ and $\xi_b(\mathbf{x}_2)$ denote the contents of the bins with index b from the histograms of $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. We say that hypothesis $\mathbf{x}_1$ is *stronger* than $\mathbf{x}_2$ if $d(\mathbf{x}_1, \mathbf{x}_2) > 0$. The second idea is to reduce the search area in the voting space using the region of interest computed from segmented clusters in the laser data. This further reduces the search space and results in a faster and more robust detection due to the scale information.

B. Second Extension to ISM: Features weight analysis

An important problem of classifying high dimensional feature vectors consist in the correct positioning of the separating hypersurfaces between negative and positive samples. The original ISM approach does not consider this problem and it just classifies the feature distribution of pedestrian feature descriptors π^+ thus, during the detection step, it uses a distance threshold T in order to match features to the codebook. In this paper we enrich the pedestrian feature set using two different informative weights in order to select and treat differently each match in the detection phase. Features found in the pedestrian silhouette and features found in the background are now both collected in the training phase. Therefore, a neighborhood analysis of each positive feature descriptor is computed considering the quantity of negative samples in a radius of distance T (the same value used in the detection step). This value called w_i^f is then normalized with respect to the cardinality of the negative set π^-:

$$w_i^f = 1 - \frac{card(neigh_i^T(\pi^-))}{card(\pi^-)} \tag{4}$$

This weight gives an information about the distinctiveness of each feature, assigning very low values to positive samples in loci where a high number of negative descriptors are found. In order to prune out *weak* feature vectors without impoverishing the learned pedestrian feature distribution, a low value of w_i^f have to be chosen. This elimination method decreases the amount of false positive matching and it can be seen as a compact way of expressing a $k - nn$ classification.

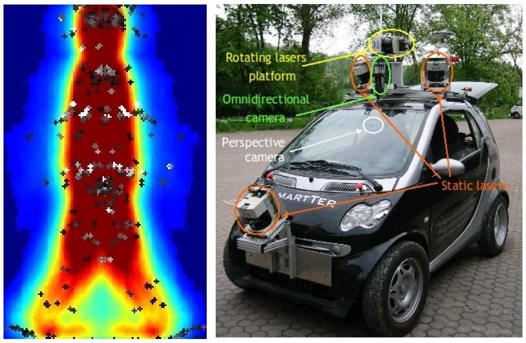

Fig. 3. **Left:** Features are weighted for their positional stability w_i^u. Features found in the trunk are more stable (white), features found in the legs are less stable due to the part motion. **Right:** Smartter vehicle platform.

Another proposed improvement in the classification method is to consider statistics in the position of the positive feature set as an informative cue of the pedestrian pose. Pedestrian features are analyzed for positional stability with respect to the object center: more the same feature is found in the same area more a high weight w_i^u is assigned. According to this weight, features found on the trunk of the pedestrian body will have high values due to its rigidness and features found on the limbs area will have a low value due to their flexibility and position change with respect to the object center. Rigid features will vote the center as a part of a rigid body keeping a fixed angle between the vector pointing to the object center and the vector parallel to the direction of its support (the direction in which the descriptor is computed to be rotation invariant). The rest of the features are classified also with their support angle and matched on the codebook during detection with a given variance in order to distinct that similar descriptors at totally different angles do not classify pedestrians (see Figure 3)

C. Third Extension to ISM: High-dimensional Nearest Neighbor Search

Another problem of the ISM-based detector is the time required to compute the matching probability $p(\mathbf{d}_i^C \mid \mathbf{d}_j^I)$. Image descriptors such as SIFT, GLOH or PCA-SIFT are very powerful (see [15] for a comparison), but they may have up to 256 dimensions. Considering that the size of the codebook can be as big as 25000, we can see that neither a linear nearest-neighbor (NN) search can be used for real-time applications or kD-trees that provide efficient NN search only for dimensions not more than 20, because the number of neighboring cells inside a given hypersphere grows exponentially with the number of dimensions.

Therefore we apply *approximate* NN search, which is defined as follows. For a given set of d-dimensional points $\mathcal{P} \subset \mathbb{R}^d$ and a given radius r, find all points $\mathbf{p} \in \mathcal{P}$ for a query point $\mathbf{q}$ so that $\|\mathbf{p} - \mathbf{q}\|_2 \leq r$ with a probability of at least $1 - \delta$. This can be implemented efficiently using locality-sensitive hashing (LSH) as proposed by [1].

D. Region of interest generation in urban environment

A common problem of ISM based methods is the tendency of generating a high quantity of false positives. In the voting stage an image feature can match several times a codebook entry and therefore it can vote for multiple object centers. Due to object symmetries, feature mismatches and scene configurations (i.e. vertical structures, complex buildings) strong false positive object hypotheses can occur in empty or unlikely areas on the image. In this paper we propose an effective and fast way to remove this kind of errors based on a distance transform computation. The idea here is that large connected ridges in the distance transform image can be safely disregarded in the detection process because they do not contain any gradient information, which is a necessary condition for the detection of a pedestrian. Two additional parameters are required here: The minimal area I_q of a ridge that can be discarded, and a safety distance I_w between a pixel and the edge that is closest to it in the image. Both of these parameters are set so that no contour of a pedestrian is included in the discarded area (in our case we use $I_q =$ and $I_w =$). This method is particularly effective in urban environments where roads and sky are often visible and contain no or little information. It consists in the following four steps (see Figure 4):

1) Compute an edge map using Canny edge detector.
2) Compute an approximate distance transform [3].
3) Cluster connected components from all points that have a distance of at least I_w to the nearest edge.
4) Discard all regions with an area that is bigger than I_q. The remaining polygonal map consitutes tha region of interest for the pedestrian detection.

The only assumption we make is that a sufficient contrast is present in the image, which is reasonable, because object detection is generally hard in low contrast images.

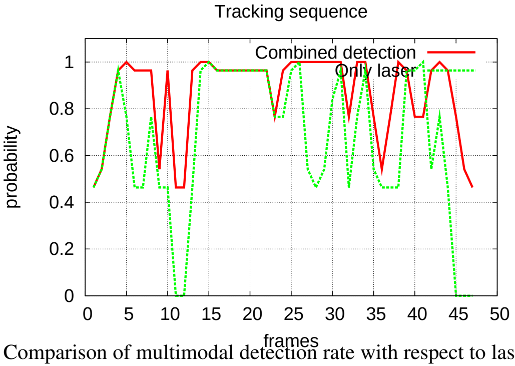

Fig. 6. Comparison of multimodal detection rate with respect to laser based people detection on a tracking sequence. The tracking follows a pedestrian and an overall higher probability is obtained with multimodal detection method than a laser detection method. A part of the graph shows that laser detection performs better in case of multiple continuous false negatives obtained from image detection but then it quickly regains confidence.

VI. Experimental Results

A. Training datasets

Our mobile platform Smartter has been equipped with an IBEO ALASKA laser scanner ($0.25deg$ resolution, $180deg$ field of view, max range up to $200m$), a rotating turntable with two SICK LMS 291-S05 lasers (3D laser) ($1.0deg$ resolution, $180deg$ field of view, max range up to $200m$), and a camera behind the windscreen (see Fig. 3).

1) Image detection: We trained our image detection algorithm using a set 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. SIFT descriptors [14] computed at Hessian-Laplace interest points are collected for the codebook building. Binary segmentation masks are used to select only features that are inside the person's shape.

2) Laser detection: We trained our laser-range detection algorithm computing several features on clustered points. Laser training datasets have been taken in different outdoor scenarios: a crowded parking lot and a university campus. The training data set is composed of 750 positive and 1675 negative samples. The resulting cascade consists of 4 stages with a total of 8 features.

B. Qualitative and quantitative results

We evaluated our extension of ISM (ISMe) on a challenging dataset. We collected two datasets in an urban environment and selected sequences in which pedestrians are walking, crossing, standing and where severe occlusions are present. Both sequences are manually annotated with bounding boxes of at least 80 pixel height and where at least half of a person body is shown. The first test set consists of 311 images containing 938 annotated pedestrians, the second consists of 171 images and 724 annotated pedestrians.

In order to show a quantitative performance several comparisons have been performed. A comparison between ISM, the proposed ISM extended (ISMe) and Haar based Adaboost (HAda) classifier is shown in the Precision-Recall graph of Fig. 5 (**top center**). Equal error rates (EER) are highlighted in each curve in order to show the performance gain. It is important to notice that at higher Recall values ISM (and HAda) shows a low precision (lots of false positives), while our method, thanks to generated ROIs and the proposed extension performs much better. HAda in general shows the limit of using boosted cascades and not robust Haar features for obtaining detection in complex backgrounds: if top level stages do not classify, the detector produces false negatives, which is often the case in a complex or occluded image frame. ISMe is significantly flatter than the other two methods and tends to the optimal upper right corner of the graph. Another comparison presented is the normalized difference in number of features processed between ISM and ISMe (Fig. 5(**top right**). ISMe works with one type of descriptor and one type of interest point, ISM usually has two or three. In average the number of descriptors to be matched and processed by ISMe is less than half than ISM. Therefore, we considered a clustered codebook and a single ROI in the image with a fixed number of features (about 150) and we activated the approximate NN in the matching step to show a speed gain of about 5 times between the two methods. Moreover, we plotted Recall over frames to show a comparison for each sample between ISMe and HAda. We can see, as we expected, AdaBoost based approach yields a very low hit rate, conversely, ISMe has a quite high true-positive rate during the entire sequence frame. Another experiment shown in the section is a comparison between ISMe and ISM in the false positive rate. Here the difference is evident and it is interesting to see that the two graphs never intersects, depiting a clear advantage of using ISMe. To quantify: ISMe, ISM and HAda obtained respectively Recall $80\%; 81\%; 78\%$ at Precision $63\%; 22\%; 0.01\%$.

We evaluated the laser classification on a data set in crowded scenes with 249 positive and 1799 negative samples.

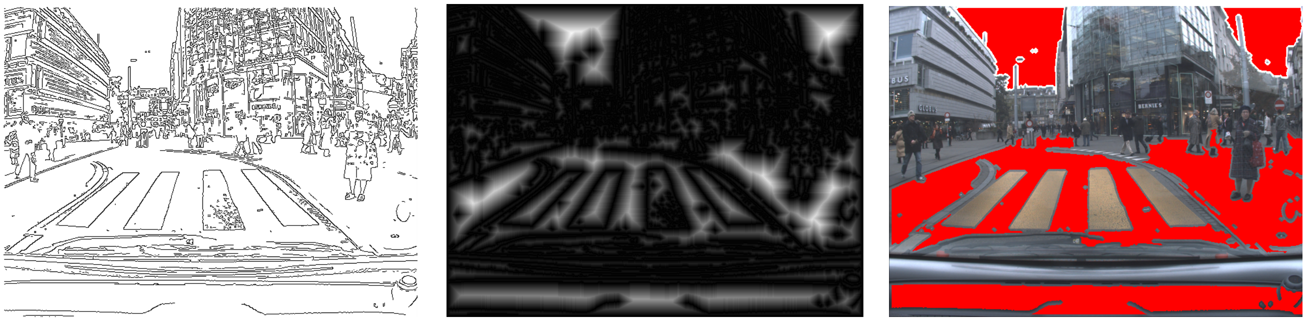

Fig. 4. Region of interest generation. Uninformative content is discarded from the image by reasoning on the distance transformed image.**Left:** Edge image (Canny). **Middle:** Approximate distance transform. **Right:** Result of the clustering in the distance transform image: areas in red are discarded

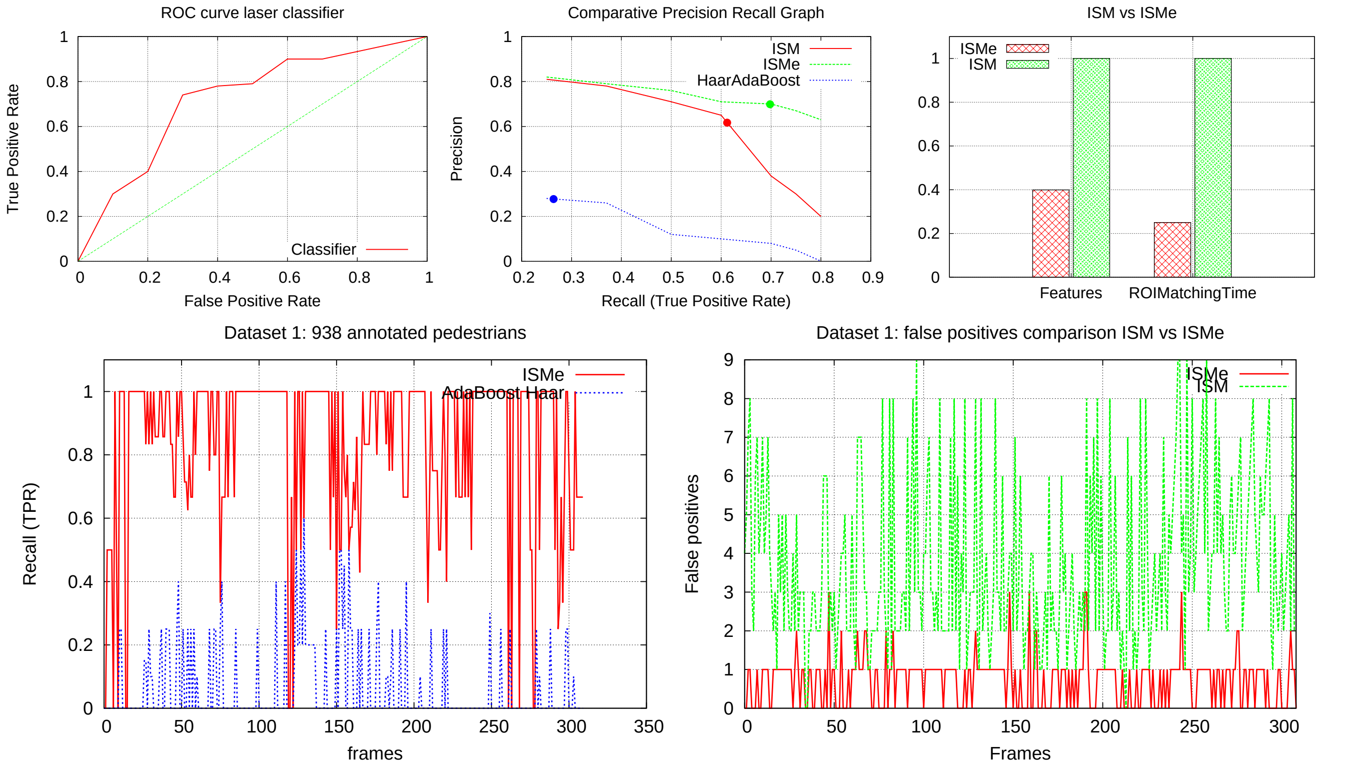

Fig. 5. **Top Left**: ROC curve for the laser classifier. AUC is ≈ 0.8.**Top Center**: Precision Recall graph to compare ISM, ISMe and Haar based Adaboost, equal error rate (EER) is respectively at $26\%; 61\%; 26\%; 69\%$. Notice the flatness of ISMe specially at high Recall values. **Top Right**: Average features quantity to match in ISM vs ISMe is shown in the left histogram; matching time is evaluated in the right histogram when approximate NN search is activated.**Lower Left** Image detection recall value on frames in dataset 1: ISMe vs Haar Adaboost cascade method. ISMe obtains an higher detection rate than the other method mainly due the distinctiveness of the features used, the detection given by a soft decision on multiple votes and the robustness against occlusion. **Lower Right** False positives in image classification evaluated for each frame of dataset 1 compared to standard ISM. Here it is clear the advantage of rescricting the voting in ROIs with the other proposed improvements.

We obtained a true positive rate (TPR) of 74.7% and a false positive rate (FPR) of 30.0% (TP:184 FN: 65 FP: 536 TN: 1273), the ROC curve is shown in Fig. 5(**top left**).

We evaluated the usefulness of the multimdal detection computing statistics of pedestrian detection at maximum range of $15m$. In order to quantify the performance of the system we considered the probability evolution of tracking a single person with both sensors and with just one 2D laser (see Fig. 6). The overall detection probability for this track increases and a smoother and more confident tracking is achieved. It is important to remark that there is a part in which the multimodal detection performs slightly worse than plain laser detection. There, a continuous false negative detection occurred in the image detector but this was quickly recovered as can be seen. We also note that many annotated pedestrians are severely occluded, and the detection task is so difficult that a performance of over 90% is far beyond the state of current computer vision systems.

Qualitative results are shown in Fig. 7. The box colors in the image describe different tracks, the size of the filled circle is proportional to the pedestrian detection confidence.

VII. Conclusions

In this paper, we presented a method to reliably detect and track people in crowded outdoor scenarios using 2D and 3D laser range data and camera images. We showed

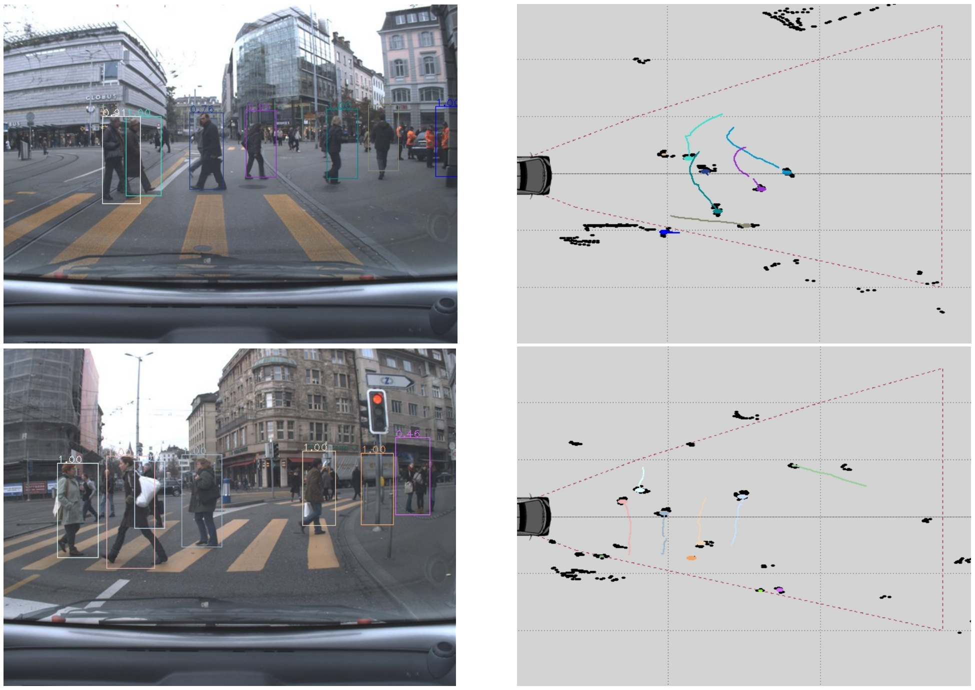

Fig. 7. Qualitative results from dataset 1 and 2 showing pedestrian crossings. The colored boxes in the image describe different tracks and probability levels; the size of the filled circle in the tracking figure is proportional to pedestrian detection confidence. It is important to notice that highly occluded pedestrians are also successfully detected and tracked.

that the detection of a person is improved by cooperatively classifying the feature vectors computed from the input data, where we made use of supervised learning techniques to obtain the classifiers. Furthermore we presented an improved version of the ISM based people detector and an EKF-based tracking algorithm to obtain the trajectories of the detected persons. Finally, we presented experimental results on real-world data that point out the usefulness of our approach.

VIII. Acknowledgments

This work was conducted and funded within the EU Integrated Projects BACS - FP6-IST-027140

References

[1] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Proc. of the Symp. on Found. of Comp. Sc.*, 2006.

[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] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics and Image Processing*, 34:344–371.

[4] D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 438–445, 2001.

[5] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, pages 66–73, 2000.

[6] 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.

[7] D. Gavrila and V. Philomin. Real-time object detection for "smart“ vehicles. In *IEEE Int. Conf. on Computer Vision (ICCV)*, 1999.

[8] D. M. Gavrila. The visual analysis of human movement: A survey. *Comp. Vis. and Image Und. (CVIU)*, 73(1):82–98, 1999.

[9] D. Hähnel, W. Burgard, and S. Thrun. Learning compact 3d models of indoor and outdoor environments with a mobile robot. *Robotics and Autonomous Systems*, 44:15–27, 2003.

[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] 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.

[13] Y. Liu, R. Emery, D. Chakrabarti, W. Burgard, and S. Thrun. Using EM to learn 3D models with mobile robots. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2001.

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journ. of Comp. Vis.*, 20:91–110, 2003.

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

[17] 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.

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

[19] L. Spinello, R. Triebel, and R. Siegwart. Multimodal people detection and tracking in crowded scenes. In *Proc. of the Twenty-Third Conference on Artificial Intelligence (AAAI)*, 2008.

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

[21] R. Triebel, W. Burgard, and F. Dellaert. Using hierachial EM to extract planes from 3d range scans. In *IEEE Int. Conf. on Rob. & Autom. (ICRA)*, 2005.

[22] 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.

[23] 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.

[24] 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.