Multimodal People Detection and Tracking in Crowded Scenes

Luciano Spinello **Rudolph Triebel** **Roland Siegwart**
Autonomous Systems Lab, ETH Zurich
Tannenstrasse 3, CLA E, 8092 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 multi-modal sensor fusion approach that utilizes 2D laser range and camera data. The data points in the laser scans are clustered using a novel graph-based method and an SVM based version of the cascaded AdaBoost classifier is trained with a set of geometrical features of these clusters. In the detection phase, the classified laser data is projected into the camera image to define a region of interest for the vision-based people detector. This detector is a fast version of the Implicit Shape Model (ISM) that learns an appearance codebook of local SIFT descriptors from a set of hand-labeled images of pedestrians and uses them in a voting scheme to vote for centers of detected people. The extension consists in a fast and detailed analysis of the spatial distribution of voters per detected person. 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. 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.

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. According to the National Highway Traffic Safety Administration report (NHTSA 2007) there were 4784 pedestrian fatalities in United States during the year 2006, which accounted for 11.6% of the total 42642 traffic related fatalities. In countries of Asia and Europe, the percentage of pedestrian accidents is even higher. The number of such accidents could be reduced if cars were equipped with systems that can automatically detect, track, and predict the motion of pedestrians. 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, 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, e.g. online path planning in crowded environments. 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 2D laser range data and CCD camera images cooperatively. This has the advantage that both *geometrical structure* and *visual appearance* information are available for a more robust detection. In this paper, we exploit this information using supervised learning techniques 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 vision data. In the detection phase, both classifiers yield likelihoods of detecting people which are fused into an overall detection probability. Finally, each detected person is tracked using multiple Extended Kalman Filters (EKF) with three different motion models and a greedy data association. This way, the filter can cope with different motion patterns for several persons simultaneously. The tracker is asynchronously updated by the detections from the laser and the camera data. The major contributions of this work are:

- An improved version of the image-based people detector by Leibe *et al.* (2005). The improvement consists in two extensions to the ISM for a reduced computation time to make the approach better suited for real-time applications.
- A tracking algorithm based on EKF with multiple motion models. The filter is asynchronously updated with the detection results from the laser and the camera.
- The integration of our multimodal people detector and the tracker into a robotic system that is employed in a real outdoor environment.

This paper is organized as follows. The next section describes previous work that is relevant for our approach. Then, we give an overview of our overall people detection and tracking system. Section 4 presents our detection method based on the 2D laser range data. Then, we introduce the Implicit Shape Model (ISM) and our extensions to the ISM. Subsequently, we explain our EKF-based tracking algorithm with a focus on the multiple motion models we

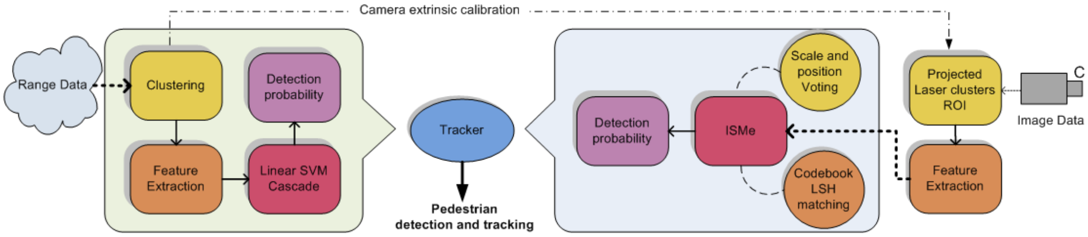

Figure 1: Overview of the individual steps of our system. See text for details.

use. Finally, we describe our experiments and conclusions.

Previous Work

Several approaches can be found in the literature to identify a person in 2D laser data including analysis of local minima (Scheutz, Mcraven, & Cserey 2004; Schulz *et al.* 2003; Topp & Christensen 2005), geometric rules (Xavier *et al.* 2005), or a maximum-likelihood estimation to detect dynamic objects (Hähnel *et al.* 2003). Most similar to our work is the approach of Arras, Mozos, & Burgard (2007) 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 (Spinello & Siegwart 2008) by using multi-dimensional features and learning them using a cascade of Support Vector Machines (SVM) instead of the AdaBoost decision stumps. In this paper, we will make use of that work and combine it with an improved appearance-based people detection and an EKF-based tracking algorithm.

In the area of image-based people detection, there mainly exist two kinds of approaches (see Gavrila (1999) for a survey). One uses the analysis of a *detection window* or *templates* (Gavrila & Philomin 1999; Viola, Jones, & Snow 2003), the other performs a *parts-based* detection (Felzenszwalb & Huttenlocher 2000; Ioffe & Forsyth 2001). Leibe, Seemann, & Schiele (2005) 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. Cui *et al.* (2005) use multiple laser scanners at foot height and a monocular camera to obtain people tracking by extracting feet and step candidates. Zivkovic & Kröse (2007) 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 (2006) 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. The RBPF tracks contours in the image based on Chamfer matching as well as point clusters in the laser scan and computes the likelihood of different prototypical shapes 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.

Several methods have been proposed to track moving objects in sequential data (see Cox (1993) for an overview). The most common ones include the joint likelihood filter (JLF), the joint probabilistic data association filter (JPDAF), and the multiple hypothesis filter (MHF). Unfortunately, the exponential complexity of these methods makes them inappropriate for real-time applications such as navigation and path planning. Cox & Miller (1995) approximate the MHF and JPDA methods by applying Murty's algorithm and demonstrate in simulations the resulting speedup for the MHF method. Rasmussen & Hager (2001) extend the JLM, JPDA, and MHF algorithms to track objects represented by complex feature combinations. Schumitsch *et al.* (2006) propose a method to reduce the complexity of MHT methods introducing the Identity Management Kalman Filter (IMKF) for entities with signature.

Overview of the method

Our system is divided into three phases: training, detection and tracking (see Fig. 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 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, from which the appearance-based classifier extracts local image descriptors and computes 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. In the following, we describe the particular steps of our system in detail.

Structure Information from Laser Data Analysis

We assume that the robot is equipped with a laser range sensor that provides 2D scan points $(\mathbf{x}_1, ..., \mathbf{x}_N)$ in the laser plane. We detect a person in a range scan by first clustering the data and then applying a boosted classifier on the clusters, which we describe as follows.

Clustering

Jump distance clustering is a widely used method for 2D laser range data in mobile robotics (see Premebida & Nunes (2005) for an overview). It is fast and simple to implement: if the Euclidean distance between two adjacent data points exceeds a given threshold, a new cluster is generated. Although this approach performs well in indoor scenarios, it gives poor results for outdoor data, because the environment is geometrically more complex and bigger distances, reflections and direct sunlight effects usually occur. This often leads to over-segmented data with many small clusters. To address this problem, we use a simple and effective technique that extends the classic jump distance method. It consists in the following steps:

1. Perform jump distance clustering with threshold ϑ. Each cluster $\mathcal{S}_i$ is defined by its left border $\mathbf{x}_i^l$, its central point $\mathbf{x}_i^c$, and its right border $\mathbf{x}_i^r$:
$$\mathcal{S}_i = \{\mathbf{x}_i^l, \mathbf{x}_i^c, \mathbf{x}_i^r\} \tag{1}$$
2. Compute a Delaunay triangulation on the centers $\mathbf{x}_i^c$.
3. Annotate each edge $\mathbf{e}_{ij} := (\mathbf{x}_i^c, \mathbf{x}_j^c)$ of the Delaunay graph with the Euclidean distance between $\mathcal{S}_i$ and $\mathcal{S}_j$.
4. Remove edges with a distance greater than ϑ and merge each remaining connected component into a new cluster.

Note that the same threshold ϑ is used twice: first to define the minimum jump distance between the end points of adjacent clusters and then to define the Euclidean distance between clusters. Experimental results showed that this reduces the cluster quantity of $25\% - 60\%$, significantly reducing overclustering. The additional computational cost due to the Delaunay triangulation and distance computation is lower compared to a full 2D agglomerative clustering approach.

Boosted Cascade of Support Vector Machines

We use AdaBoost (Freund & Schapire 1997) to classify the clustered laser data into the classes "person" and "no person". AdaBoost creates a strong classifier from a set of weak classifiers. Viola & Jones (2002) further improved this approach by ordering the weak classifiers in a degenerate decision tree which they call an *attentional cascade*. This reduces the computation time significantly. We apply this method, but we use support vector machines (SVMs), in particular c-SVMs with linear kernel (Boser, Guyon, & Vapnik 1992), instead of the standard decision stumps based on thresholding. 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. Before applying the SVMs, we normalize the input data in order to avoid numerical problems caused by large attribute values. The parameters c of the c-SVMs where obtained from a local search where the classification results where evaluated using 5-fold cross validation.

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

$$p(\pi \mid \mathbf{f}) = \sum_{i=1}^{L} w_i h_i(\mathbf{f}) \tag{2}$$

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 (2) for each feature vector $\mathbf{f}$ in the test data set. In our implementation, we compute the features $\mathbf{f}$ of a cluster $\mathcal{S}$ as described in our previous work (Spinello & Siegwart 2008).

Appearance Information from Image Data Analysis

Our image-based people detector is mostly inspired by the work of Leibe, Seemann, & Schiele (2005) on scale-invariant Implicit Shape Models (ISM). An ISM is a generative model for object detection and has been applied to a variety of object categories including cars, motorbikes, animals and pedestrians. 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 $\{(\Delta x_{ij}, \Delta y_{ij})\}$ and scale factors $\{s_{ij}\}$ 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. We can think of this as a sample-based representation of a spatial distribution $p(\pi, \hat{\mathbf{x}} \mid \mathbf{d}_i^C, \mathbf{x}_i)$ for each $\mathbf{d}_i^C$ at a given image location $\mathbf{x}_i = (x_i, y_i)$ where $\hat{\mathbf{x}} = (\hat{x}, \hat{y})$ denotes the center of the detected person. 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. To compute the likelihood to detect a person at location $\bar{\mathbf{x}}$ we use the following marginalization:

$$p(\pi, \bar{\mathbf{x}} \mid \mathbf{x}_j^I, \mathbf{d}_j^I) = \sum_{i=1}^{K} p(\pi, \bar{\mathbf{x}} \mid \mathbf{d}_i^C, \mathbf{x}_j^I) p(\mathbf{d}_i^C \mid \mathbf{d}_j^I) \quad (3)$$

This defines the weight of the vote that is cast by each descriptor $\mathbf{d}_j^I$ at location $\mathbf{x}_j^I$ for a particular occurrence of a person at position $\bar{\mathbf{x}}$. The overall detection probability is then the sum over all votes:

$$p(\pi, \bar{\mathbf{x}} \mid \mathbf{g}^I) = \sum_{j=1}^{M} p(\pi, \bar{\mathbf{x}} \mid \mathbf{x}_j^I, \mathbf{d}_j^I) \quad (4)$$

where $\mathbf{g}^I = (\mathbf{x}_1^I, \ldots, \mathbf{x}_M^I, \mathbf{d}_1^I, \ldots, \mathbf{d}_M^I)$. With the sample-based representation, we can find the $\bar{\mathbf{x}}$ that maximizes (4) by a maxima search using a variable-bandwidth mean shift balloon density estimator (Comaniciu, Ramesh, & Meer 2001) in the 3D voting space.

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. Leibe, Seemann, & Schiele (2005) 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 about 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 according to (3) we also compute the angle α

$$\alpha(\mathbf{x}_j^I, \mathbf{x}) = \arctan2(y_j^I - y, x_j^I - x) \quad (5)$$

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

$$d(\mathbf{x}_1, \mathbf{x}_2) := \sum_{b=1}^{B} \xi_b(\mathbf{x}_1) - \xi_b(\mathbf{x}_2), \quad (6)$$

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.

Second 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 effective (see Mikolajczyk & Schmid (2005) 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 a linear nearest-neighbor (NN) search can not be used for real-time applications. A potential alternative would be the use of kD-trees, but these provide efficient NN search only for dimensions not more than around 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 Andoni & Indyk (2006).

Tracking Pedestrians

So far, we described how pedestrians can be detected in 2D laser range data and in camera images. The result of these detectors is an estimate of the position of a person at a given time frame. However, for many applications it is required to also have information about the *kinematics* of the person, e.g. provided by a motion vector. This can be achieved by tracking the position of the person and predicting the future motion based on the observations from the previous time frames. A key issue for a people tracking algorithm is the definition of the motion model. Pedestrians are not constrained to a particular kind of motion and they can abruptly change their motion direction at any time. To address this problem, we use the following motion models for each tracked person:

1. **Brownian motion:** This accounts for sudden motion changes.
2. **Constant speed:** The person does not change direction or speed.
3. **Smooth turning:** The forward speed is constant and a we fit a second order polynomial into the last 10 positions of the pedestrian using least mean square fitting (LMS).

In each time step, the tracker needs to solve the data association problem that consist in finding a mapping between observations and tracked objects. In our system, we use a greedy approach to do the data association, which is performed in two steps: In the first step we choose the motion model whose prediction has the smallest distance to the closest observation. In the second step, we consider the person with the longest tracking history and assign to it the observation that is closest to it and still inside a 3σ ellipse from the last position. For the distance computation we use the Mahalanobis metric. Then we assign the observation that is closest to the person with the second-longest history and so on. If this process ends up with unassociated observations, a

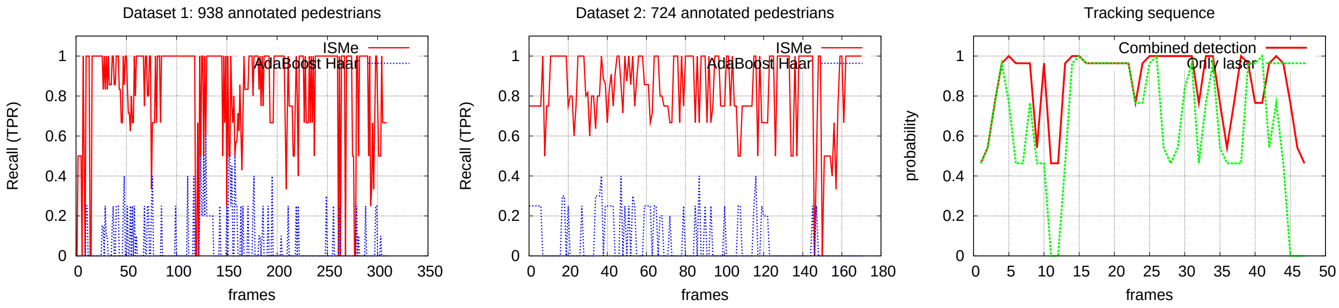

Figure 2: **Left** and **Center:** Image detection recall value and frames in dataset 1 and 2: ISMe vs Haar Adaboost cascade method. ISMe yields a higher detection rate than Haar/Adaboost mainly due to the distinctiveness of the features used, the detection based on a soft decision on multiple votes, and the robustness against occlusions. **Right:** Comparison of multimodal and laser-only based people detection on a tracking sequence. The tracker follows a pedestrian and a higher overall probability is obtained with the multimodal detection method compared to the laser-only detection. A part of the graph shows that the laser detection performs better in case of multiple continuous false negatives in the image detection, but then the algorithm quickly regains confidence.

new track is created. In the case that no assignment is found for a tracked person, the corresponding track is updated only using the motion prediction based on all three motion models. This is done until a new observation can be assigned, but at most for 0.5 seconds, afterwards the track is removed.

Experimental Results

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

Training datasets

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 (Lowe 2003) 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.

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.

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's 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 a comparison is performed between the classic Haar based AdaBoost pedestrian detection and our detector (see Figure 2 left and center). We can see that the AdaBoost based approach yields a very low hit rate (on average less that 50%) on both datasets due to the low robustness of Haar features and the concept of the cascade. If top level stages do not classify, the detector produces false negatives, which is often the case in a complex or occluded image frame. Conversely, ISMe has a quite high true-positive rate (TPR) during the entire frame sequence. Recall and precision rates have been computed in order to verify the role of false positives for ISMe. For both datasets the computed recall value is similar and comparably high (82% and 81%). Similar results are obtained for the precision ($\approx 61\%$). ISMe has also been compared with an unconstrained implementation of ISM (maximum strength center selection and no image ROI constraint) and the resulting precision was half of the ISMe precision value.

In order to quantify the laser classification, a test data set in crowded scenes composed of 249 positive and 1799 negative samples data was prepared. We obtained a true positive rate (TPR) of 64.7% and a false positive rate (FPR) of 30.0% (FP:161 FN: 88 FP: 536 TN: 1273). To test the usefulness of using a multimodal detection algorithm a single person was tracked, and a comparison with a laser-only detection is shown in the right plot of Fig. 2. 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 occured in the image detector, but this was quickly recovered as can be seen. We also note that many pedestrians were severely occluded, and that the detection task is so difficult that a performance of over 90% is far beyond the state of current computer vision systems.

Qualitative results from two frames are shown in Fig. 3. The box colors in the image correspond to different tracks, and the size of the filled circles is proportional to the pedestrian detection confidence. Another experiment has been performed to evaluate the time advantage of using an LSH approach during codebook matching with respect to linear

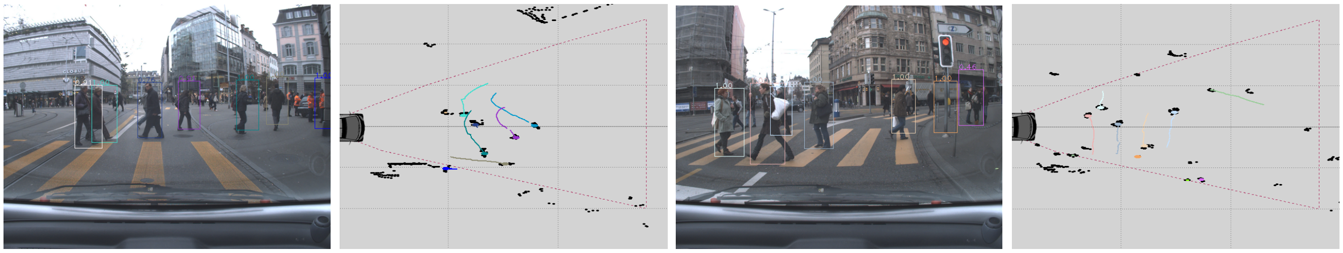

Figure 3: Qualitative results from dataset 1 and 2 showing pedestrian crossings. The colored boxes in the image describe different tracks and probability levels; the size of the filled circle in the tracking figure is proportional to the confidence of the pedestrian detection. It is important to notice that highly occluded pedestrians are also successfully detected and tracked.

neighbor search. A clustered codebook has been produced and tested by matching a random test image extracted from one of the two sequences with the codebook. LSH-based NN search resulted 12 times faster than the linear approach.

Conclusions

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

Acknowledgment

This work has been supported by the EC under contract number FP6-IST-027140-BACS.

References

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

Arras, K. O.; Mozos, Ó. M.; and Burgard, W. 2007. Using boosted features for the detection of people in 2d range data. In *IEEE Int. Conf. on Rob. & Autom. (ICRA)*.

Boser, B. E.; Guyon, I.; and Vapnik, V. 1992. A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, 144–152.

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

Cox, I. J., and Miller, M. L. 1995. On finding ranked assignments with application to multitarget tracking and motion correspondence. *Trans. Aerospace and Electronic Systems* 31:486–489.

Cox, I. J. 1993. A review of statistical data association for motion correspondence. *Int. Journ. of Computer Vision* 10(1):53–66.

Cui, J.; Zha, H.; Zhao, H.; and Shibasaki. 2005. Tracking multiple people using laser and vision. In *IEEE Int. Conf. on Intell. Rob. and Sys. (IROS)*, 2116–2121.

Felzenszwalb, P., and Huttenlocher, D. 2000. Efficient matching of pictorial structures. In *IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR)*, 66–73.

Freund, Y., and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55(1):119–139.

Gavrila, D., and Philomin, V. 1999. Real-time object detection for smart vehicles. In *IEEE Int. Conf. on Comp. Vision (ICCV)*.

Gavrila, D. M. 1999. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding (CVIU)* 73(1):82–98.

Hähnel, D.; Triebel, R.; Burgard, W.; and Thrun, S. 2003. Map building with mobile robots in dynamic environments. In *IEEE Int. Conf. on Rob. & Autom. (ICRA)*.

Ioffe, S., and Forsyth, D. A. 2001. Probabilistic methods for finding people. *Int. Journ. of Computer Vision* 43(1):45–68.

Leibe, B.; Seemann, E.; and Schiele, B. 2005. Pedestrian detection in crowded scenes. In *IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR)*, 878–885.

Lowe, D. 2003. Distinctive image features from scale-invariant keypoints. *Int. Journ. of Computer Vision* 20:91–110.

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

NHTSA. 2007. 2006 traffic safety annual assessment – a preview. Traffic Safety Facts, National Center for Statistics and Analysis. http://www-nrd.nhtsa.dot.gov/Pubs/810791.PDF.

Premebida, C., and Nunes, U. 2005. Segmentation and geometric primitives extraction from 2d laser range data for mobile robot applications. In *Robotica 2005 - Scientific meeting of the 5th National Robotics Festival*.

Rasmussen, C., and Hager, G. D. 2001. Probabilistic data association methods for tracking complex visual objects. *IEEE Trans. on Pattern Analysis & Machine Intelligence* 23(6):560–576.

Scheutz; Mcraven; and Cserey. 2004. Fast, reliable, adaptive, bimodal people tracking for indoor environments. In *IEEE Int. Conf. on Intell. Rob. and Sys. (IROS)*.

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

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

Schumitsch, B.; Thrun, S.; Guibas, L.; and Olukotun, K. 2006. The identity management Kalman filter (IMKF). In *Robotics: Science and Systems (RSS)*.

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

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

Viola, P., and Jones, M. 2002. Robust real-time object detection. *Int. Journ. of Computer Vision*.

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

Xavier, J.; Pacheco, M.; Castro, D.; Ruano, A.; and Nunes, U. 2005. Fast line, arc/circle and leg detection from laser scan data in a player driver. In *IEEE Int. Conf. on Rob. & Autom. (ICRA)*, 3930–3935.

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