Multimodal feedback fusion of laser, image and temporal information

Ivan Huerta
DPDCE, University IUAV
Santa Croce 1957
Venice, Italy 30135
huertacasado@iuav.it

Gonzalo Ferrer
Institut de Robòttica i
Informàtica Industrial
Llorens i Artigas 4-6
Barcelona, Spain 08028
gferrer@iri.upc.edu

Fernando Herrero
Institut de Robòttica i
Informàtica Industrial
Llorens i Artigas 4-6
Barcelona, Spain 08028
fherrero@iri.upc.edu

Andrea Prati
DPDCE, University IUAV
Santa Croce 1957
Venice, Italy 30135
aprati@iuav.it

Alberto Sanfeliu
Institut de Robòttica i
Informàtica Industrial
Llorens i Artigas 4-6
Barcelona, Spain 08028
sanfeliu@iri.upc.edu

ABSTRACT

In the present paper, we propose a highly accurate and robust people detector, which works well under highly variant and uncertain conditions, such as occlusions, false positives and false detections. These adverse conditions, which initially motivated this research, occur when a robotic platform navigates in an urban environment, and although the scope is originally within the robotics field, the authors believe that our contributions can be extended to other fields. To this end, we propose a multimodal information fusion consisting of laser and monocular camera information. Laser information is modelled using a set of weak classifiers (Adaboost) to detect people. Camera information is processed by using HOG descriptors to classify person/non person based on a linear SVM. A multi-hypothesis tracker trails the position and velocity of each of the targets, providing temporal information to the fusion, allowing recovery of detections even when the laser segmentation fails. Experimental results show that our feedback-based system outperforms previous state-of-the-art methods in performance and accuracy, and that near real-time detection performance can be achieved.

Keywords

1. INTRODUCTION

Human beings are so accustomed to navigating in crowded environments, such as busy streets or shopping malls, that they do not even realize the extreme difficulty that executing such tasks entails. Under the scope of robotics, we aim to obtain a successful perception system that permits us to enhance the current mobile robotic navigation paradigm, and to this end, a robust and fast human detector system is mandatory.

Given a robotic platform like a two-wheeled robot ([12] and [15]), the challenge is being capable of building a system that perceives and predicts human behaviour during navigation tasks. However, the basis for the high-level interpretation of observed patterns of human motion requires detecting where the human being is. Given the nature of the project, where a highly uncertain environment is constantly sensed and the response of the system depends on human behaviour, we propose a feedback-based system that integrates a laser rangefinder, monocular camera and temporal information to detect human beings. As we will demonstrate later, the fusion of these sensors provides a tremendously robust performance, even under occlusion conditions due to the temporal information, while achieving a high level of accuracy.

Although fusion systems have been thoroughly investigated, it is still a wide and open problem, many of the works on robot navigation do not obtain the accuracy required for a safe navigation at human-standard velocities, resulting in inaccurate and extremely slow systems.

Our feedback-based approach outperforms state of the art methods, since it is able to detect people even when laser or image detections are not possible, thanks to the fusion of laser, camera and temporal information given by the feedback from the multi-hypothesis tracker. Moreover, it is able to work in nearly real-time due to the nature of each selected detector. All the modules work under the Robot Operating System (ROS) [11], which facilitates the implementation and the re-usability of the code enormously. The approach has also been widely tested in real urban scenarios.

The remainder of the paper is organised as follows. The state of the art in the field of fusion detection will be discussed in section 2. In section 3, the algorithm for laser and image people detection, along with their fusion, the tracking process and the temporal information are outlined. Finally,

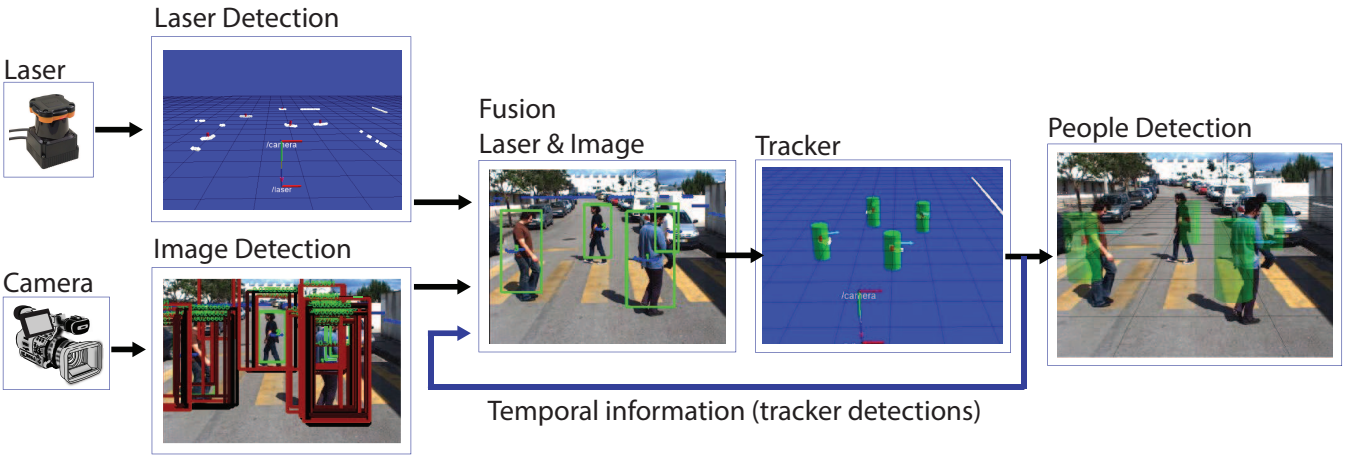

Figure 1: Overview of the method

we present experimental results in section 4 and concluding remarks in section 5.

2. STATE OF THE ART

The fusion of range and vision sensors has been addressed classically in two ways:

1. A laser scanner segmentation method is used to find the most likely ROIs; later these ROIs are projected into the image, and finally classified by an image classifying system.
2. Considering laser and vision independently and merging both in a classifier or at feature level.

Accordingly, Szarvas *et al.* [14] used a laser sensor to obtain the ROIs, which are projected into the image and classified by a convolutional neural network (CNN). Broggi *et al.* [3] proposed a method to determine the ROIs using a laser sensor, this information is used later in a Haar-like feature/adaboost classifier to name it. Mählisch *et al.* [8] and Doullard *et al.* [5] proposed a spatio-temporal alignment to integrate laser and camera sensors. They use the features from both sensors to feed a Bayesian classifier and a conditional random field (CRF) respectively to detect cars. Spinello *et al.* [13] propose a cooperative fusion of independent sensor detections. The fusion is done by data association of a tracking system in each sensor space, after having used Adaboost and CRF in the laser space and implicit shape model (ISM) in the image space. Another approach that fuses independently laser and camera sensors is presented by Oliviera *et al.* [9] which consists of a semantic fusion framework, which uses the contextual information to integrate both sensors. However, the approach is far from being capable of working in real time. Due to the constraints of our method to work on a robot, we need to achieve a trade-off between speed and accuracy. Our approach uses laser and vision independently, and merges both at a classifier level using image and depth information at the same time.

3. METHODOLOGY

In the present paper we propose a people detector system which scheme is depicted in Fig. 1. The system firstly consists of: a laser detection module which uses a boosting technique for detection; an image detection module, mainly based on HOGs descriptors and a linear SVM for classifying person/non-person; a Fusion module where spatial, depth and temporal information from image and laser detections is fused taking advantage of the three cues at the same time. Later, a tracking module is implemented using a multi-hypothesis particle filter approach. Finally, by establishing a feedback connection, as depicted in Fig. 1, we allow our fusion module to work with Laser-Image-Tracking (temporal) information, thereby allowing our approach to be able to recover detections even when there are partial occlusions or the laser segmentation fails.

3.1 Laser people detector

Our implementation of the laser detector is fundamentally based on the approach of Arras *et al.* [2] with little variation. The objective is to detect people making use of two dimensional range scans with a supervised learning classifier based on simple features extracted from groups of neighboring beams. Range measurements that correspond to humans have certain geometrical properties such as size, circularity or convexity, so the idea is to determine a set of meaningful scalar features that quantify these properties and use them in a supervised learning technique (boosting) to select the best features and thresholds for the classifier. However, achieving an accurate detection is very difficult because problems such as physical variation of people's appearance, viewpoint changes, background clutter or occlusions prevent the laser from gathering enough information.

The laser observation consists of a set of beams represented in polar coordinates (angle, length) projected to Cartesian coordinates (x,y) by the method proposed in [17]. These sets of points (x,y) are grouped in segments based on a jump distance condition, which evaluates the distance to adjacent points. For each of these segments, fourteen geometric features are determined: number of points, standard deviation, mean, average deviation from median, minimum and maximum distance between the distance to previous segment and distance to next segment, linearity, circularity,

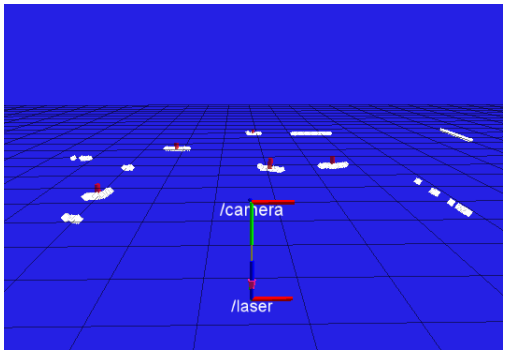

Figure 2: Laser detections. White lines are the laser scans, and red squares are the laser detections.

Figure 3: Image detector. The score of the detections are represented by their bounding box color: green represents very high score, black medium score, and red very low score.

radius, boundary length, boundary regularity, mean curvature, mean angular difference and width of pair with nearest segment. Segments with less than three points will be discarded.

The employed boosting algorithm is AdaBoost. Each example in the training data is a segment with a label associated, $+1$ or -1. In the training phase a strong clasifier is created by combining a set of weak classifiers. The weak classifiers are decision trees of one level depth (stumps) that depend on single-valued features and use a threshold and sign. On each iteration weights are normalized, a weak classifier is used to classify the examples for each feature, and a classification error is calculated. Weights are updated increasing the ones corresponding to the examples that were incorrectly classified by the last weak classifier. The final strong classifier is a weighted majority vote of the best weak classifiers.

As a result the laser detection procedure gives a vector with all the detections $L_d = [x, y, z, s]$, where x, y, z are the central position of the laser detection in Cartesian coordinates, and s is the quality (score) of the laser detection. As we have commented before, one of the major drawbacks of laser detections is their failure when there is a lack of information about the object to detect. One direct consequence of this problem is the high number of false positive detections. However, the image detector could complement this lack of data. Therefore, the fusion between the laser and image detection (section 3.3) will decide which are the correct detections. Fig. 2 shows an image with laser detections, where white lines are the laser scans, and the red squares are the people laser detections centered at laser segments.

3.2 Image People Detection

The object detector selected for the image classification is based on Dalal *et al.* [4] approach, which use Histogram of Oriented Gradient (HOG) features with Support Vector Machines (SVM).

HOG have been largely used as robust visual descriptors in many computer vision applications related to object detection and recognition, due to their expressiveness, fast computation, compactness, and invariance to misalignment and monotonic illumination changes. The features are built by partitioning an input image or region into connected cells, counting occurrences of gradient edge orientations within each cell into a histogram with evenly spread channels, and concatenating the cell histograms in a single array. Depending on whether the sign of the gradient is taken into account, gradient orientations are sampled from the $[0; \pi]$ or $[-\pi; \pi]$ domain. The energy of each cell histogram is typically normalized in order to compensate for illumination variations. Once Histogram of Oriented Gradient feature vectors are extracted, they are fed to a linear SVM for object/non-object classification. The detection window is scanned across the image at all positions and scales to detect object instances in the output pyramid.

Usually, the object detector is able to detect almost all the objects in the scene, however sometimes the detections have such a low score that are not considered correct. If we want to detect these false negative detections, then we will detect a lot of false positives. The idea is that instead of setting a threshold to determine if a detection is correct or not, we make use of the spatial, depth and temporal information to make this decision. Therefore, non-maximum suppression such as grouping or deleting based on lower scores, relative position, or based on mean shift is not employed in our approach, to make sure that partially occluded detections are not deleted just because they are close to other detections.

Fig. 3 shows image people detections, where the score of the detections are represented by their bounding box color. Green bounding boxes represent positive score detections, black bounding boxes show score detections around 0, and red bounding boxes represent negative score detections. Usually, the object detector only takes into consideration the detections which are positives, whereas we will consider all the detections in the fusion stage. As a result the image detection gives a vector with all the detections $I_d = [x, y, w, h, s]$, where $[x, y]$ and $[w, h]$ are the upper-left corner coordinates and the size of the bounding box detection respectively, and s is the score provided by the image detector.

Other descriptors might have been used as well, such as rectified Haar wavelets used by Papageorgiou *et al.* [10], or directly using edge images, similarly to Gavrila *et al.* [6]. Also, instead of using a SVM classifier, another object detector can be used such as Adaboost employed by Viola *et*

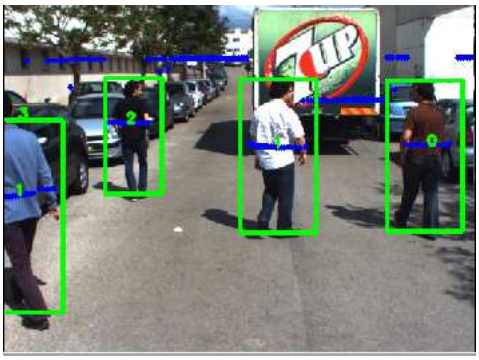

Figure 4: Fusion laser and image detector. Blue lines are laser scans, numbers are input detections (laser and tracking inputs), and green bounding boxes are image detections.

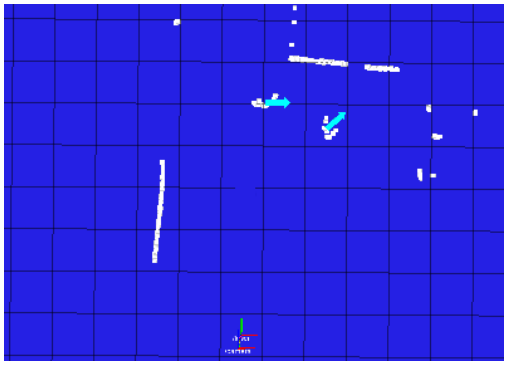

Figure 5: Tracker detections. White lines are laser scans, and cyan arrows are tracker outputs.

al. [16].

3.3 Fusion Laser and Image Detection

The data provided by the camera are the perfect complement to the laser scanner detector due to its different sensing nature. While detections by the camera usually fail under certain environmental conditions, they contain more information about people than detections made by the laser scanner. In contrast, the detections made by the laser, which are independent of environmental conditions, usually produce more false positives because of the limited data used in the process.

In order to measure the accuracy of the detections, we proceed as follows. Firstly, image detections are filtered based on the bounding box size $I_{d,w,h}$ and laser depth $L_{d,z}$ of the detections, thereby taking into consideration the 3D geometry of the scene. Secondly, the 2D-spatial relationship between the image detections (I_d) and the input detections (N_d) are taken into account. The accuracy of each input detection is measured with respect to the image detections. The input detections (N_d) in the first iteration are only the laser detections (L_d): these detections also include those with a very low score. In this way, all the segments found as possible detections are tested. In the next iterations the input detections (N_d) also contain the temporal information provided by the tracker (T_d), thereby allowing the approach to be robust against laser detection failures.

The probability for each of the input detections is calculated based on the image detections score and the spatial relationship between the input detection and the bounding box image detection. Accordingly, we calculate the distance between the bounding box detection and the input detection normalized by the distance between the corner of the bounding box with the possible position of the input detection in the image:

$$dist(I_d, N_d) = \frac{\sqrt{(Ipos_x - N_{d,y})^2 + (Ipos_y - N_{d,y})^2}}{\sqrt{(Ipos_x - I_{d,y})^2 + (Ipos_y - I_{d,y})^2}} \quad (1)$$

where $Ipos_{x,y}$ is the predicted input detection position in the image detection I_d calculated as follows:

$$\begin{aligned} Ipos_x &= I_{d,x} + I_{d,w} * Lpos_x \\ Ipos_y &= I_{d,y} + I_{d,h} * Lpos_y \end{aligned} \quad (2)$$

Lpos is the relative position of the laser detection with respect to the bounding box detections. *Lpos* is learned in the training process or is the position between the laser device respect to the ground plane of the image. Therefore, *Ipos* is the position of the *Lpos* in the current image detection I_d.

Finally, the detection score is calculated as follows:

$$S(d) = \left(\sum_{I_d \in N_d} I'_{d,s} * (1 - dist\,(I_d, N_d)) \right) * \frac{|I_d \in N_d|}{|N_d|} \quad (3)$$

where $I'_{d,s}$ is the unity-based normalized score of the image detection:

$$I'_{d,s} = \frac{I_{d,s} - mI_d}{mI_d - nI_d} \quad (4)$$

$I_{d,s}$ corresponds to the score for each image detection, and nI_d and mI_d are the *min* and *max* score for all the image detections, respectively. $|I_d \in N_d|$ denotes the number of image detections for the input detection, and $|N_d|$ the number of input detections.

An example of fusion between image and laser detections can be seen in Fig. 4, where blue lines are laser scans, numbers are the input detections (could be laser or tracker data), and green bounding boxes correspond to image detections. In Fig. 4, input detection number 3 (localized at the upper-left of bounding box 1) is discarded by the image detector.

3.4 People tracking

The people tracking follows a straightforward implementation of the work done at the Freiburg University by Luber [7] and Arras [1]. That is a multi-hypotheses tracker, using linear propagation that can handle occlusions, crossings and loss of targets, at a relative low error. However, instead of using a Kalman filter, we use a particle filter, whose particles consist of position and velocity for each of the targets in four dimensions ($T_d = [x, y, z, v]$). In Fig. 5 a snapshot of the tracker output is depicted, plotted as cyan arrows.

3.5 Fusion of laser and image detection with temporal information

Figure 6: Example of detection using fusion with temporal information. Left Image, image without the temporal information used as feedback. Right image, image with the temporal information used as feedback. The right image shows that the person has been correctly detected.

Temporal information provided by the tracking T_d feeds back the vision system in order to avoid losing true positive detections, otherwise not detected by the laser detector, by adding temporal consistency to the approach. Therefore, the inputs to the fusion detection module (sec. 3.3) are the image detections I_d, the laser information L_d, and the temporal information provided by the tracking detections T_d as depicted in Fig. 1.

Temporal consistency is important for improving the detection because the detections where the laser fails or where the laser beams are occluded can be recovered. Furthermore, image detections can correct the problem of incorrect tracking detections. Fig. 6 shows an example where a person is detected thanks to the feedback from the temporal information. The left image is without the temporal information used as feedback whereas the right image is with our fusion approach with the temporal information.

4. EXPERIMENTAL RESULTS

All the modules described so far have been implemented[1] using the Robot Operational System (ROS) and were designed to work with our robot [15] for more details). However, we have used the ISR-UC-imglidar-sync datasets[2] published by Oliveira *et al.* [9], in order to validate the system performance. The laser scanner was mounted at a height of 0.9 m to detect pedestrians at waist level. The system sensing ranges from 2 up to 20 m. The laser scanner was set with an aperture of 100° and angular resolution of 0.25°, while the camera forms a field of view of 45°. For a description of the dataset please refer to [9] paper.

As shown in table 1, our feedback approach obtains a better detection rate than [9]. Our feedback-based approach gets 84.13% HR (hit rate) at FAPF=0.5 (false alarm per frame) while [9] gets 80.8% HR (obtained from [9]). Moreover, as was expected, our approach using feedback information obtains better performance than without using feedback.

The approach has been widely tested in real conditions as it has been implemented and currently works online in a two-wheeled robot. The entire method works nearly in real time at 7 fps with a Intel Core i7 3930K, 3,2Ghz hexa core.

[1]All code implemented is available at http://www.ros.org/wiki/iri-ros-pkg
[2]http://www.visionandbrain.com/datasets.html

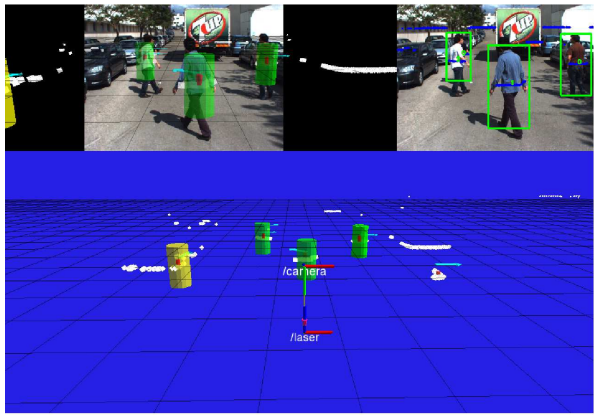

Figure 7: People detection modules output. Green cylinders and big green bounding boxes are the people detections results by our fusion approach.

Approach	Hit rate
Semantic Fusion [9]	80.8
Our approach (no tracker feedback)	79.2
Our approach with tracker feedback	84.1

Table 1: Hit rate for pedestrian detection at FAPF=0.5 (False alarm per frame)

An image of the results is depicted in Fig. 7. Where the output of laser detections (red little cylinders), the fusion detections (green cylinders), and the detections not confirmed by the image detector because they are outside the image (yellow cylinders) can be seen. In addition the tracking output is reported with cyan arrows, and the white and blue lines represent the laser scans. All of them are plotted in a 3D space (bottom image) with their corresponding projection into the image plane (upper left corner). The fusion detections with their bounding boxes are in the upper right corner in the image. Fig. 8 shows significant frames using our feedback-fusion approach from the ISR-UC-imglidar-sync dataset.

5. CONCLUSION

A novel approach has been presented in this paper, which is able to detect and track people from a mobile robot platform. The presented feedback-fusion method combines laser, image and temporal information, taking advantage of the three cues at the same time. The tracking information feeds back the image detection, fusing this information with laser detections, thereby allowing the recovery of detections even when the laser and image segmentation fail. The proposed architecture represents a step forward over state-of-the-art methods, as has been shown in the experimental results. Furthermore, the method has been tested in real situations, is implemented in ROS and currently works in a two-wheeled robot performing nearly in real-time. Future works should deal with image detection because it is the bottleneck which prevents system faster performance. Moreover, the tracking should be improved to work in crowded environments.

Figure 8: People detection modules output using our feedback-fusion approach from ISR-UC-imglidar-sync database. Green cylinders and big green bounding boxes are the people detections results by our fusion approach, respectively. See text for details.

Acknowledgments.
The authors would like to thank to Christina Zitello for English editing and Joan Perez for his collaboration. This work has been partially funded by the European project CargoANTs (FP7-SST-2013- 605598) and by the Spanish CICYT project DPI2013-42458-P.

6. REFERENCES

[1] K. Arras, S. Grzonka, M. Luber, and W. Burgard. Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. In *ICRA*. IEEE, may 2008.

[2] K. Arras, O. Mozos, and W. Burgard. Using boosted features for the detection of people in 2d range data. In *ICRA*. IEEE, April 2007.

[3] A. Broggi, P. Cerri, S. Ghidoni, P. Grisleri, and H. Gi. A new approach to urban pedestrian detection for automatic braking. *Trans. Intell. Transp. Syst.*, 10, 2009.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, 2005.

[5] B. Douillard, D. Fox, and F. Ramos. A spatio-temporal probabilistic model for multi-sensor object recognition. In *IROS*, pages 2402–2408, 2007.

[6] D. M. Gavrila, J. Giebel, , and S. Munder. Vision-based pedestrian detection: the protector+ system. In *IEEE IV'04*, 2004.

[7] M. Luber, G. Diego Tipaldi, and K. Arras. Place-dependent people tracking. *IEEE Int. J. Robot Res.*, 30(3):280, Jan. 2011.

[8] M. Mählisch, R. Schweiger, W. Ritter, and K. Dietmayer. Sensorfusion using spatio-temporal aligned video and lidar for improved vehicle detection. In *IEEE IV'06*, pages 424–429, 2006.

[9] L. Oliveira, U. Nunes, P. Peixoto, M. Silva, and F. Moita. Semantic fusion of laser and vision in pedestrian detection. *PR*, 43:3648–âĂŞ3659, 2010.

[10] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 38:15–33, 2000.

[11] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng. ROS: an open-source Robot Operating System. In *ICRA Workshop on Open Source Software*, 2009.

[12] A. Sanfeliu and J. Andrade-cetto. Ubiquitous Networking Robotics in Urban Settings. In *IEEE/RSJ IROS Workshop on Network Robot Systems*, 2006.

[13] L. Spinello, R. Triebel, and R. Siegwart. A trained system for multimodal perception in urban environments. In *ICRA Workshop on Safe Navigation in Open and Dynamic Environment*, 2009.

[14] M. Szarvas, U. Sakai, and J. Ogata. Real-time pedestrian detection using lidar and convolutional neural networks. In *IEEE IV'06*, 2006.

[15] E. Trulls, A. Corominas Murtra, J. Pèrez-Ibarz, G. Ferrer, D. Vasquez, J. M. Mirats-Tur, and A. Sanfeliu. Autonomous navigation for mobile service robots in urban pedestrian environments. *Journal of Field Robotics*, 28(3):329–354, 2011.

[16] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *ICCV*, 2003.

[17] Q. Zhang and R. Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *IROS*, pages 2301–2306, 2004.