*Workshop on Robotic Co-workers 4.0: Human Safety and Comfort in Human-Robot Interactive Social Environments*
*2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, Madrid, Spain*

# Fast People Detection for Mobile Robot based on Multiple Sensors and Convolutional Neural Networks

Ali Youssef, Daniele Nardi and María T. Lázaro

***Abstract*— Developing automatic, fast and robust people detection approach can help mobile robot platforms to navigate safely with socially acceptable behavior in human populated environments. In this paper we present a novel approach for combining data from 2D laser scanners and RGB images in one fast human detection module. The fusion of leg detectors with deep learning classifiers enhances detection performance based only on detectors, and enabling the use of deep learning techniques in robot applications with the presence of limited hardware (e.g., no GPU).**

## I. Introduction

People detection by moving platforms has been widely studied in recent years due to advances in applications where robots operate in human populated environments. Ensuring human safety and comfort in such environments requires effective and robust people perception modules. Common sensors used in people detection are RGB and depth cameras and 2D and 3D laser range finders (LRF). Due to the presence of environmental changes (e.g., illumination), the variety of postures of people (e,g., walking, standing), their non standardized size and their dynamism, people detection by moving robotic platform with each type of sensor data has its advantages and drawbacks. Pedestrian detection with 2D LRF have been addressed mainly by using machine learning classifiers [1] and [2] and combination of geometric and statical features [3]. Although LRF provides wide field of view, high data rate and invariance to illumination changes, the robustness of the detectors still require to be improved due to false positive detections.

In computer vision, people detectors were developed to handle full-body [4], upper-body [5] and partial human part detection [6]. Recently, the success of convolutional neural networks (CNN) in achieving the top results of object detection and image classification [7] has attracted the research to explore the ability of achieving robust people detection, which can deal with a more realistic and complex environment. Due to the limited computational power available on most robots and the computational cost of previous computer vision approaches (i.e., CNN), adapted techniques based on data acquired by different sensors have been studied to allow the use of CNN in real time robotic applications. In [5] people detection approaches are used based on different type of sensors. By reducing the search space and using simple network architecture [8], CNN have achieved good object detection and facilitated the use of CNN in real time scenarios.

In this paper we present a novel approach for combining leg detection by 2D LRF with a supervised image classification based on deep CNN in RGB images. The approach is developed to be used on our social mobile robot DIAGO[1] to run social human-robot interaction and cognitive robotics experiments. A laser based leg detector [2] is used to provide the regions where proper people are located. A coarse-to-fine detection techniques is followed by clustering laser scan points and ROIs extraction. Consequently, pruning of false positives caused by the leg detector in a later classification process, and reduction in search space in the images have been gained. Based on the person's distance from the robot, two CNN based classifiers (upper-body and full-body) are used for ROIs validation. The main contributions of this paper are: i) an integration of 2D LRF and RGB camera for robust people detection; ii) deep CNN based upper-body and full-body person classifiers. We present computation time and classification performance quantitative results obtained by the proposed integrated approach on indoor data set acquired by DIAGO.

## II. Related Work

Several works have been presented for people detection based on 2D LRF [2] [1]. The works aimed to use laser data as an independent measurement of the environmental conditions with wide field of view. Leg features are extracted as single blobs based on geometric information or by clustering nearby points that match assumption of being legs. Those blobs are used in detection-by-tracking presented people or classifiers based on AdaBoost and random forest classifiers into person or not-person. However, depending only on laser sensor for people detection has some disadvantages: i) limited information can be extracted, consequently higher level tasks (e.g., reasoning and re-identify) will not be used with such approaches; ii) false positive detections in cluttered environment which requires other assumptions to prune them and complex features extraction to be processed which increase the computational cost and reduce the generalization properties of detectors.

With the rise of robot employment in domestic and public environments, the demand of accurate and fast human

Ali Youssef, Daniele Nardi and María T. Lázaro are with Dipartimento di Ingegneria Informatica Automatica e Gestionale "Antonio Ruberti", Sapienza University of Rome, Italy. {youssef, nardi,mtlazaro}@diag.uniroma1.it

[1]https://sites.google.com/a/dis.uniroma1.it/diago/

detection is increasing. Hand engineered features combined with machine learning classifier in [4] [9] have advantages in feature extraction and classification and disadvantages regarding computational cost. However, the state-of-the-art of object detection has been achieved by deep learning models [10] [11] [12]. Region-based convolutional neural network (R-CNN) [11] provides bounding box of the detected object based selective search and regions are merged depending on similarity metrics and variety of color spaces. Faster region-based [12] convolutional network (faster R-CNN) used the region proposal network (RPN) to reduce the computational cost caused by the selective search. Localizing objects in the image with its class probability prediction all at once in a single evaluation can be achieved by you only look once (YOLO) [13] and single-shot detector (SSD) [14]. Along with significant information extracted by the CNN, the computational cost of the detector has paramount importance in robotic applications where limited power of CPU cores or even small percentage of the CPU cycle are available for the detection task enabling the robot to perform other tasks like navigation and human interaction. The main challenges to employ recent object detection methods in robotic applications can be abstracted as: i) computational cost; ii) human recognition and localization in real world coordinates which requires further processing than only image processing. Therefore, it seems useful to explore the viability of multi-sensor fusion solution to the problems of using CNN based object detection in robotic applications. In [1] ROIs are validated using HOG descriptor with SVM classifier. In [8], preprocessing procedure is used based on color segmentation enabling the use of simple architecture of CNN classifier. The depth segmentation is used instead of laser reducing the search space and providing the proposed ROIs to predict the object's class. In [15] depth-template matching approach is used based on a trained classifier to detect the upper body in close range. The work in [16] projected the depth ROIs into RGB image space to be validated using fine-tuning of pre-trained CNN model. In [5], fusion of information acquired by detection procedure is used in tracking system. The detection is achieved in parallel by independent laser and camera sensors and can be used only with filtering assumptions based on ground plane estimation and GPU for far range person detection. Here, our aim is to combine open source fast leg detector with less computational load classification methods, in order to speed up the use of CNN for pedestrian detection even in presence of a limited hardware. We believe that the adoption of deep learning techniques adds an additional level of robustness to the robot vision system and can lead to the use of context information for higher-level tasks (e.g., safe navigation and human-robot spatial interaction).

## III. Proposed Approach

During the last few years, different researches proposed solutions for using deep learning for pedestrian detection without providing the ability to merge it with a classification step [17]. Our method adopts the same approaches [1], i.e., detection followed by classification, with a difference that the detection process relies on CNN validation with simple architectures that were trained from scratch, have better detection performance and can be used on limited hardware. The functional architecture is shown in Fig. 1, where both laser and RGB sensing are leveraged in one people detection pipeline. As a possible approach to employ deep learning techniques on limited hardware, we propose to have a pre-processing step that allow to reduce the amount of input information (i.e., by reducing the search space) to be classified by simple CNN architecture developed for person detection. This reduction is provided by leg detection process and ROIs extraction functions in the pipeline.

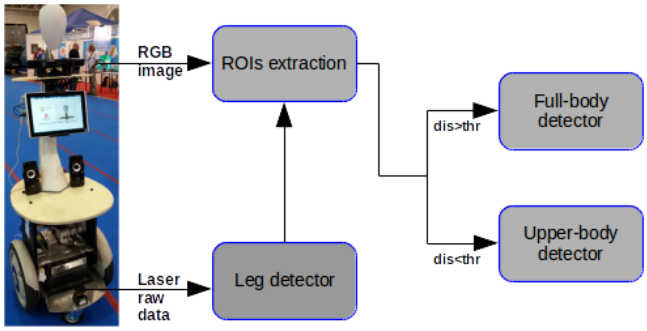


Fig. 1. Proposed approach.

### A. Leg detection

For leg detection we use the standard ROS component `leg_detector`[2]. This module takes as input raw laser data and uses a machine-learning-trained classifier to provide a set of 3D points $\mathcal{L} = \{\mathbf{p}_i^l\}, \mathbf{p}_i^l = (l_x, l_y, l_z) \in \Re_3$ in the laser frame corresponding to possible positions of detected legs of people. The results provided by this module alone are susceptible to false positives, especially in dynamic and cluttered environments, which motivates the use of novel computer vision techniques in a verification step presented in the next sections.

### B. ROIs extraction

The set of detected leg points $\mathcal{L}$ represented in laser frame are provided to the ROIs extraction procedure. This step aims at first to project each $\mathbf{p}_i^L$ into image plane. Given $\mathbf{T}_c^R$, the pose of the camera with respect to robot frame and $\mathbf{T}_L^R$ the pose of laser with respect to the robot, it is possible to obtain the leg point $\mathbf{p}_i^C$ in camera frame by applying on each leg point $\mathbf{p}_i^L$ the following transformation:

$$\mathbf{T}_L^C = (\mathbf{T}_C^R)^{-1} \cdot \mathbf{T}_L^R$$
$$\mathbf{p}_i^C = \mathbf{T}_L^C \cdot \mathbf{p}_i^L$$

[2]http://wiki.ros.org/leg_detector

By using the intrinsic calibration, point in camera frame $\mathbf{p}_i^c$ can be projected into image plane to get pixel coordinates $\mathbf{u} = (u, v)$ of each detected leg. Based on camera matrix $\mathbf{K}$ we can obtain a leg point in the image plane coordinates as following:

$$\mathbf{p}_i^I = \mathbf{K} \cdot \mathbf{p}_i^C$$

$\mathbf{p}_i^I = (x_I, y_I, z_I)$ is projected leg point into image plane. To get the pixel coordinates:

$$\big(u\ v\big)^T = \big(\frac{x_I}{z_I}\ \frac{y_I}{z_I}\big)^T$$

Detected legs that are projected out of camera field of view are ignored. Based on average distance and dimension of an adult person, we calculated the ROI's height as in [1] and we set the width to be half of calculated height. Moreover, cropped ROIs are separated based on their distance from robot into upper-body and full-body ROIs.

### C. CNN classifiers

Two binary CNN based classifiers (upper-body and full-body classifiers) are placed at the end of the pipeline to obtain a validation of ROIs. Due to the distortion caused by resizing full-body images to fit the common CNNs input, we developed full-body classifier that keeps the size ratio of the input image. With the assumption that full-body size is $128 \times 64$ pixels with ratio 2:1 of height to width, the network structure presents $4$ convolutional, 2 max-pooling and 2 fully connected layers followed by softmax with loss layer. Relu is used as an activation function. For convolutional layers we used kernel of size $3 \times 3$ and 32 filters. Max-pooling with a kernel of size $3 \times 3$ is used after the first and fourth convolutional layers, while we aim to learn transformed weights by convolutional operation in other layers and reducing the size of model. The work in [17] provides good experimental study on different CNN structure for person detection. For upper-body classifier, we use an input size of $64 \times 64$ pixels. The network structure consists of $4$ convolutional layers, 3 average pooling layers and one fully connected layer followed by softmax with loss and sigmoid function has been used as an activation function. We used stochastic gradient descent as an optimizer in both classifiers.

## IV. Experimental Results

This section presents the evaluation of the proposed approach. Along the experiments we aim at demonstrating the effectiveness of the proposed approach in terms of computational load and generalization properties of CNN model. The pipeline is developed using ROS[3] and the Caffe[4] framework for the classifiers. We run the experiments on a Intel(R) Core(TM) i7-3770 CPU@3.40GHz $\times 8$ with 16GB RAM.

### A. Platform

DIAGO mobile robot (Fig. 1) on board sensors were used for data acquisition. DIAGO is equipped with an *Hokuyo UTM-30LX*[5] LRF on the bottom part of the robot and a *Microsoft Kinect*[6] on the top, which constitutes a typical configuration in social robots. A correct extrinsic calibration by DIAGO's sensors provide $\mathbf{T}_L^R$ and $\mathbf{T}_C^R$ which enables transformation between laser frame and camera frame. Moreover, intrinsic camera parameters represented by $\mathbf{K}$ matrix are obtained to project points into the image plane (i.e., compute corresponding pixels of 3D world points in the image plane).

### B. Data set

In order to train the binary classifiers, we use two types of data set for each. For full-body classifier, we used images obtained from MIT pedestrian images [18] and INRIA person data set [4]. For upper-body classifier, images were obtained from [19]. To avoid overfitting problem, we use Dropout [20] and batch normalization layers and data set augmentation. Each class has $5000$ images in total (e.e., full-body, unknown object and upper-body). We divide them into the ratio 7:2:1 for training, validation and test, respectively. In order to prove the generalization of our model, we use another set of data obtained by DIAGO robot to test the classifiers within the detection pipeline. We recorded different realistic indoor data sets and we evaluate the classifiers on almost $1000$ cropped images provided by the ROIs extraction step. We set the distance threshold to $2m$ to enable the use of proper classifier and we consider far people within $15m$.

### C. Evaluation

Confusion matrices comparison becomes difficult to analyze when different type of data set is used. We use four well-known metrics in the evaluation, which are defined as following:

$$Sensitivity = \frac{TP}{TP + FN}$$
$$Specificity = \frac{TN}{TN + FP}$$
$$Precision = \frac{TP}{TP + FP}$$
$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$TP$ is the number of true positive ROIs classified as person), $FP$ is the number of false positive ROIs, $TN$ is the number of true negative ROIs, and $FN$ is the number of false negative ROIs. Qualitative results of the proposed detection pipeline are illustrated in Fig. 2. First row shows the projection of laser scan points into image plane, then the ROIs extraction provides batches to be classified. The second row in Fig. 2 illustrates detection performance in different indoor scenarios.

[3]http://www.ros.org/
[4]http://caffe.berkeleyvision.org/
[5]https://www.hokuyo-aut.jp
[6]https://developer.microsoft.com/en-us/windows/kinect

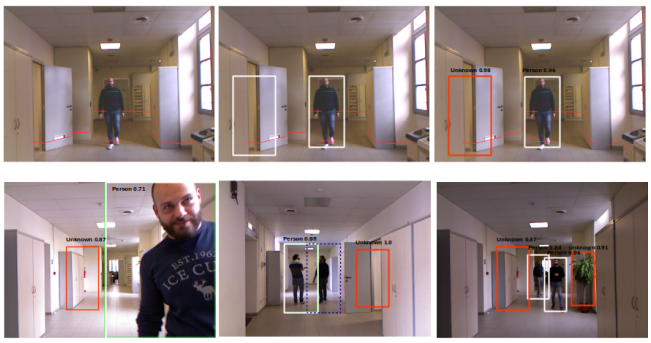

Fig. 2. Qualitative results. The first row shows the processing steps of the proposed pipeline, where red dots represents the laser scan points projected into image plane, white rectangles are TP ROIs and red rectangles are TN negative samples with corrected prediction by CNN classifiers. The second row shows only the output prediction of the pipeline in various scenarios where white rectangles are TP ROIs representing person, red rectangles are unknown object and blue rectangle is FN sample missed by leg detector.

The approach has been evaluated separately at first. Several CNN structures have been tested for better trade-off between computational load and classification accuracy. Table I shows quantitative results of both classifiers with the the computational time of processing each ROI using **CPU**. The laser based leg detector evaluation can be found at [2]. Moreover, we tried to measure the leg detection performance on our own data set. The leg detector has good

TABLE I

QUANTITATIVE RESULTS OF PROPOSED CNN CLASSIFIERS.

| Approach | Sensitivity | Specificity | Precision | Accuracy | Time (s) |
|---|---|---|---|---|---|
| Upper-body | 0.94 | 0.97 | 0.98 | 0.94 | 0.016 |
| Full-body | 0.95 | 0.97 | 0.98 | 0.96 | 0.011 |

performance in terms of TP samples, but it provides a lot of FP samples especially in indoor environments. This was one of our work's motivation (FP reduction). This improvement will enable use the proposed approach successfully in higher level tasks such as tracking and reasoning. Table II shows the detection performance of full pipeline and the advantages gained by combining the leg detector with CNN classifiers. It should be noticed that leg detection evaluation has been achieved in terms of *sensitivity* and *precision* due to the difficulties of measuring the TN term.

TABLE II

QUANTITATIVE RESULTS TO COMPARE THE PERFORMANCE OF PROPOSED DETECTION PIPELINE TO ONLY LEG DETECTION APPROACH

| Approach | Sensitivity | Specificity | Precision | Accuracy | Time (s) |
|---|---|---|---|---|---|
| CNN-leg | 0.94 | 0.94 | 0.97 | 0.94 | 0.02 |
| leg detector | 0.78 | - | 0.7 | - | 0.004 |

## V. CONCLUSIONS

In this paper, we presented a novel approach to combine two people detectors based on different types of sensors in one human detection pipeline. Quantitative results show the improvement with respect to leg-only detection, together with our generalized, fast and simple CNN classifiers which allows its use in a broad of range of robotic applications. Future directions of our work will consider the use of the proposed detection pipeline for further human intention analysis for human-robot interaction applications.

## REFERENCES

[1] E. P. Fotiadis, M. Garzón, and A. Barrientos, "Human detection from a mobile robot using fusion of laser and vision information," *Sensors*, vol. 13, no. 9, pp. 11 603–11 635, 2013.
[2] K. O. Arras, O. M. Mozos, and W. Burgard, "Using boosted features for the detection of people in 2d range data," in *IEEE Int. Conf. on Robotics and Automation*, 2007, pp. 3402–3407.
[3] L. Spinello and R. Siegwart, "Human detection using multimodal and multidimensional features," in *IEEE Int. Conf. on Robotics and Automation*. IEEE, 2008, pp. 3264–3269.
[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 886–893.
[5] C. Dondrup, N. Bellotto, F. Jovan, M. Hanheide *et al.*, "Real-time multisensor people tracking for human-robot spatial interaction," 2015.
[6] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.
[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[8] D. Albani, A. Youssef, V. Suriani, D. Nardi, and D. D. Bloisi, "A deep learning approach for object recognition with nao soccer robots," in *Robot World Cup*. Springer, 2016, pp. 392–403.
[9] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
[10] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.
[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2014.
[12] R. Girshick, "Fast r-cnn," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 1440–1448.
[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.
[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conf. on Computer Vision*. Springer, 2016, pp. 21–37.
[15] O. H. Jafari, D. Mitzel, and B. Leibe, "Real-time rgb-d based people detection and tracking for mobile robots and head-worn cameras," in *IEEE Int. Conf. on Robotics and Automation*, 2014, pp. 5636–5643.
[16] A. Broad and B. Argall, "Geometry-based region proposals for real-time robot detection of tabletop objects," *arXiv preprint arXiv:1703.04665*, 2017.
[17] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 4073–4082.
[18] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 1997, pp. 193–199.
[19] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.
[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.