The IIIA30 Mobile Robot Object Recognition Dataset

Arnau Ramisa[†], David Aldavert[§], Shrihari Vasudevan[‡], Ricardo Toledo[§] and Ramon Lopez de Mantaras[†]

† Artificial Intelligence Research Institute (IIIA-CSIC), Campus UAB, Bellaterra, E-08193, Spain.
Email: aramisa@iiia.csic.es and mantaras@iiia.csic.es
§ Computer Vision Center, Campus UAB, Bellaterra, E-08193, Spain.
Email: aldavert@cvc.uab.cat and ricard@cvc.uab.cat
‡ Australian Center for Field Robotics, the University of Sydney, NSW 2006, Australia
Email: shrihari.vasudevan@ieee.org

***Abstract*—Object perception is a key feature in order to make mobile robots able to perform high-level tasks. However, research aimed at addressing the constraints and limitations encountered in a mobile robotics scenario, like low image resolution, motion blur or tight computational constraints, is still very scarce. In order to facilitate future research in this direction, in this work we present an object detection and recognition dataset acquired using a mobile robotic platform. As a baseline for the dataset, we evaluated the cascade of weak classifiers object detection method from Viola and Jones.**

I. Introduction

Currently there is a big push towards semantics and higher level cognitive capabilities in robotics research. One central requirement towards these capabilities is to be able to identify higher level features like objects, doors, etc. For example, in [1], the authors investigate underlying representations of spatial cognition for autonomous robots. Although not specifically addressed in that work, object perception is an essential component that the authors reported to be the most limiting factor.

Although different modalities of perception (e.g. laser range-finder, color camera, haptics) can be used, in this work we focus on passive vision, as it is interesting for several reasons like an affordable cost, passive and low power consuming, compatibility with human environments or richness of perceived information.

Recently several methods have been quite successful in particular instances of the problem, such as detecting frontal faces or cars, or in datasets that concentrate on a particular issue (e.g. classification in the Caltech-101 [2] dataset). However in more challenging datasets like the detection competition of the Pascal VOC 2007 [3] the methods presented achieved a lower average precision. This low performance is not surprising, since object recognition in real scenes is one of the most challenging problems in computer vision [4]. The visual appearance of objects can change enormously due to different viewpoints, occlusions, illumination variations or sensor noise. Furthermore, objects are not presented alone to the vision system, but they are immersed in an environment with other elements, which clutter the scene and make recognition more complicated. In a mobile robotics scenario a new challenge is added to the list: computational complexity. In a dynamic world, information about the objects in the scene can become obsolete even before it is ready to be used if the recognition algorithm is not fast enough.

Despite the importance of the problem, we are not aware of any publicly available dataset where the particular problems of mobile robotics are well represented. To help improve this situation, we have created the IIIA30 dataset, which contains several sequences acquired navigating with a mobile robot, as well as manually generated bounding box annotations of 29 different objects.

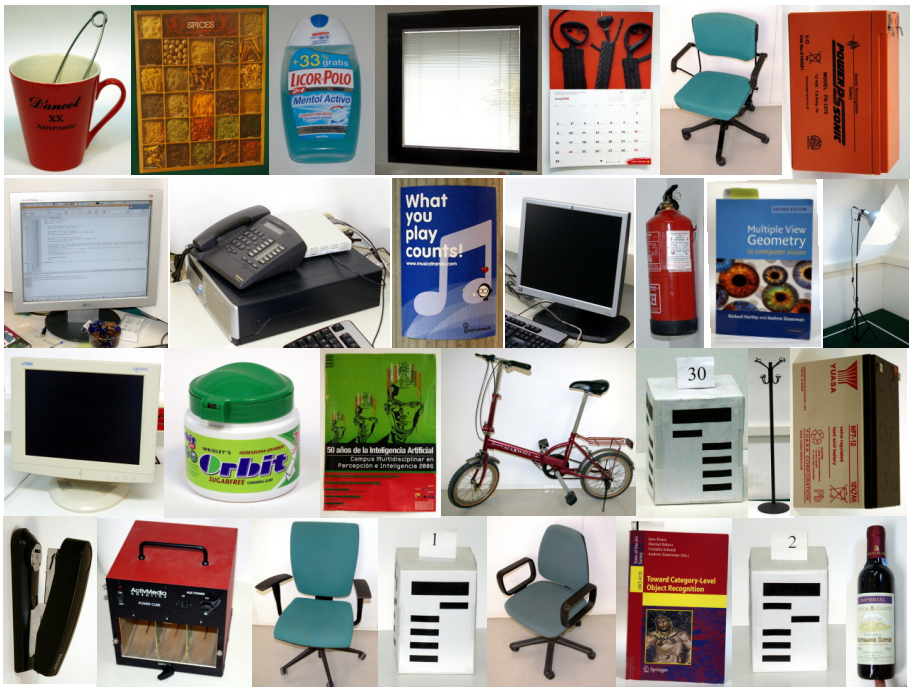
Moreover, the problems encountered in mobile robotics and embodied in this dataset are very similar to those found in mobile computing, a currently very relevant area of research where low processing power, limited storage space and bad image quality are the rule rather than the exception.

The rest of the paper is divided as follows. First, the IIIA30 dataset and the performance metrics we recommend for it are described in Section II. Next, the dataset is evaluated using the well known cascade of weak classifiers method from Viola and Jones in Section III. Finally, in Section IV, the conclusions are presented.

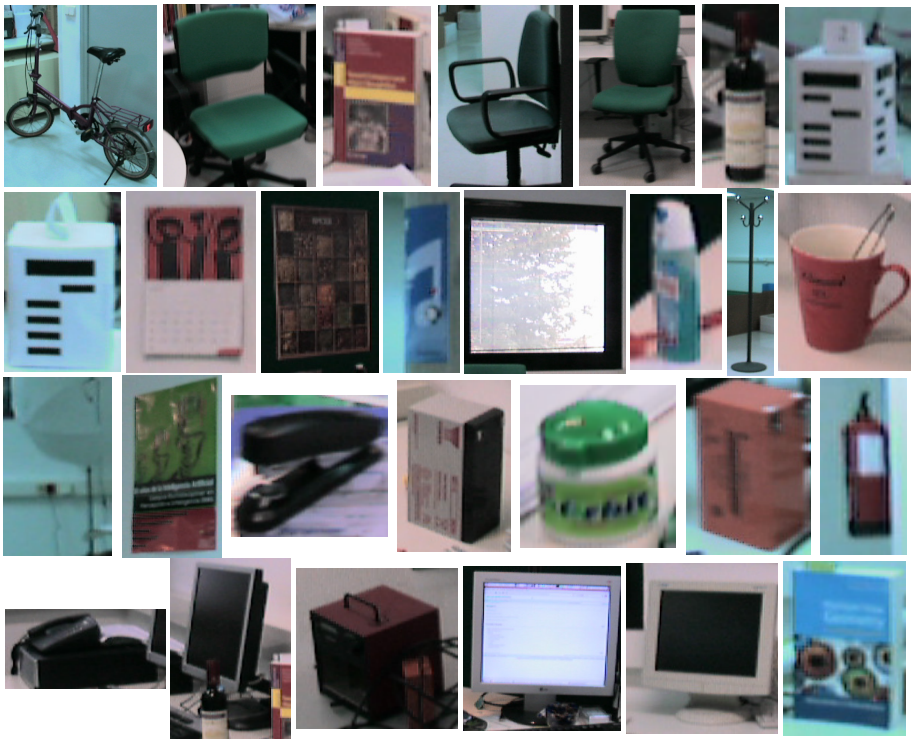
II. Dataset and Performance Metrics

We have created the IIIA30 dataset[1], that consists of three sequences (IIIA30-1 to IIIA30-3) of different length acquired by our mobile robot while navigating at approximately 50 cm/s in a laboratory type environment, and approximately twenty good quality images for training taken with a standard digital camera. The camera mounted in the robot is a Sony DFW-VL500 and the image size is standard VGA resolution (i.e. 640×480 pixels). In Figure 1 the robotic platform used can be seen. The environment has not been modified in any way and the object instances in the test images are affected by lightning changes, blur caused by the motion of the robot, occlusion and large viewpoint and scale changes.

[1] http://www.iiia.csic.es/~aramisa/iiia30.html



(a)



(b)

Fig. 2. (a) Training images for the IIIA30 dataset. (b) Cropped instances of objects from the test images.

Since the objective was to deal with a non-trivial multiclass problem, a total of 30 categories (29 objects and background) that appear in the sequences have been considered. In order to evaluate the influence of different training datasets, twenty good quality training images and a video were taken with a standard digital camera for each considered object category. The objects have a large range of sizes, and cover a wide range of appearance characteristics: some are textured and flat, like the posters, while others are textureless and only defined by its shape. Figure 2.a shows the training images for all the object categories, and 2.b shows some cropped object instances from the test images. Each occurrence of an object in the video sequences has been manually annotated in each frame to construct the ground truth, along with its particular image characteristics (e.g. blurred, occluded...).

In order to evaluate the performance of the methods we recommend several standard metrics that are briefly explained in the following lines. Precision is defined as the ratio of true positives among all the positively labeled examples, and reflects how accurate our classifier is.

$$Pre = \frac{TruePositives}{FalsePositives + TruePositives} \quad (1)$$

Recall measures the percentage of true positives that our classifier has been able to label as such. Namely,

$$Rec = \frac{TruePositives}{FalseNegatives + TruePositives} \quad (2)$$

When it is equally important to perform well in both metrics, we also considered the f–Measure metric:

$$f - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (3)$$

This measure assigns a single score to an operating point of our classifier weighting equally precision and recall, and is also known as f_1–measure or balanced f–score. If the costs of a false positive and a false negative are asymetric, the general f–measure can be used by adjusting the β parameter:

$$f_g - measure = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (4)$$

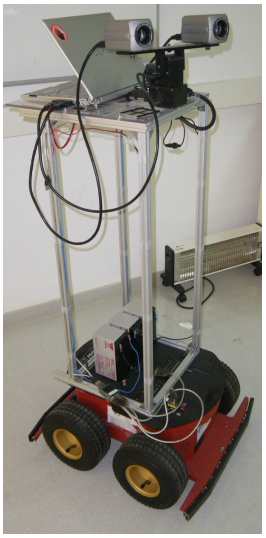


Fig. 1. Robotic platform used in the experiments.

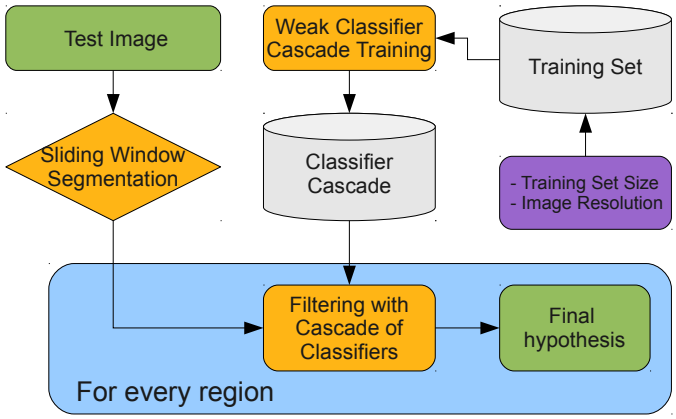


Fig. 3. Diagram of the Viola and Jones Cascade of Weak Classifiers method, with tests shown as purple boxes. Orange boxes refer to steps of the method and green to input/output of the algorithm.

In the object detection experiments, we have used the Pascal VOC object detection criterion [3] to determine if a given detection is a false or a true positive. In brief, to consider an object as a true positive, the bounding boxes of the ground truth and the detected instance must have a ratio of overlap equal or greater than 50% according to the following equation:

$$\frac{BB_{gt} \cap BB_{detected}}{BB_{gt} \cup BB_{detected}} \geq 0.5 \quad (5)$$

where BB_{gt} and $BB_{detected}$ stand for the ground truth and detected object bounding box respectively. For objects marked as occluded only the visible part has been annotated in the ground truth. Since the type of annotation is not compatible with the output of algorithms the estimate the pose of the whole object from the visible part (like the SIFT object recognition method [11]), for the case of objects marked as occluded, we have modified the above formula in the following way:

$$\frac{BB_{gt} \cap BB_{detected}}{BB_{gt}} \geq 0.5 \quad (6)$$

As can be seen in the previous equation, it is only required that the detected object bounding box overlaps 50% of the ground truth bounding box. Another option to deal with the occluded objects problem could be modifying the object detection algorithm to restrict the predicted bounding box to the part of the object that is believed to be visible, although that would arise other theoretical and technical difficulties.

III. Baseline Results

The cascade of weak classifiers proposed by Viola and Jones [5] is a commonly used object recognition method because of its good performance and low computational cost. A diagram of the steps of the method and the tests conducted can be seen in Figure 3. This method constructs a cascade of simple classifiers (i.e. simple Haar-like features in a certain position inside a bounding box) using a learning algorithm based on AdaBoost. Speed was of primary importance to the authors of [5], and therefore every step of the algorithm was designed with efficiency in mind. The method uses rectangular Haar-like features as input from the image computed using Integral Images, which makes it a constant time operation regardless of the scale or type of feature. Then, a learning process that selects the most discriminative features constructs a cascade where each node is a filter that evaluates the presence of a single Haar-like feature with a given scale at a certain position in the selected region. The most discriminative filters are selected to be in the first stages of the cascade to discard windows not having the object of interest as soon as possible. At classification time, the image is explored using sliding windows. However, thanks to the cascade structure of the classifier it is only at interesting areas where processor time is really spent.

Notwithstanding its well known advantages, this approach suffers from significant limitations. The most important one being the amount of data required to train a competent classifier for a given class. Usually hundreds of positive and negative examples are required (e.g. in [6] 5000 positive examples, derived using random transformations from 1000 original training images, and 3000 negative examples where used for the task of frontal face recognition). Another known drawback is that a fixed aspect ratio of the objects is assumed with this method, that may not be constant for certain classes of objects (e.g. cars). Another drawback is the difficulty of generalizing the approach above 10 objects at a time [7]. Finally, the tolerance of the method to changes in the point of view is limited to about 20°. In spite of these limitations, the Viola and Jones object detector has had remarkable success and is widely used, especially for the tasks of car and frontal face detection.

Since the publication of the original work by Viola and Jones, many improvements to the method have appeared, for example to address the case of multi-view object recognition [8], [9]. In this work the original method has been evaluated using a publicly available implementation[2].

Training Set Size and Image Quality: As previously mentioned, one of the most important limitations of the Viola and Jones object recognition method is the amount and quality of the training data. In this work we have evaluated three different training sets. The first one consists of images extracted from the ground truth bounding boxes from test sequences IIIA30-2 and IIIA30-3. The second one consists of 20 good quality training images per object type, and additional synthetic views automatically generated from these images. Finally, the third training set is a mix between good quality images extracted from videos recorded with a digital camera (for 21 objects, between 700 and 1200 manually segmented images per object), and a single training image plus 1000 new synthetic views (for 8 objects).

[2]We have used the implementation that comes with the OpenCV library: http://opencv.willowgarage.com/wiki/

Object	Recall	Prec	Object	Recall	Prec
Grey battery	0.0	0.0	Monitor 2	0.14	0.14
Red battery	0.28	0.02	Monitor 3	0.03	0.01
Bicycle	0.46	0.07	Orbit box	0.03	0.01
Ponce book	0.0	0.0	Dentifrice	0.0	0.0
Hartley book	0.03	0.01	Poster CMPI	0.17	0.15
Calendar	0.19	0.01	Phone	0.0	0.0
Chair 1	0.11	0.22	Poster Mystrands	0.36	0.27
Chair 2	0.71	0.05	Poster spices	0.46	0.06
Chair 3	0.0	0.0	Rack	0.0	0.0
Charger	0.0	0.0	Red cup	0.0	0.0
Cube 1	0.0	0.0	Stapler	0.03	0.01
Cube 2	0.0	0.0	Umbrella	0.03	0.02
Cube 3	0.0	0.0	Window	0.36	0.2
Extinguisher	0.0	0.0	Wine bottle	0.0	0.0
Monitor 1	0.0	0.0			

TABLE I
Recall and precision values obtained training the Viola & Jones object detector using images extracted from the IIIA30-3 sequence and evaluating in sequences IIIA30-1 and IIIA30-2.

Object	Recall	Prec	Object	Recall	Prec
Grey battery	0.01	0.02	Monitor 2	0.41	0.20
Red battery	0.08	0.04	Monitor 3	0.40	0.18
Bicycle	0.01	0.10	Orbit box	0.10	0.16
Ponce book	0.08	0.31	Dentifrice	0.01	0.03
Hartley book	0.04	0.08	Poster CMPI	0.10	0.05
Calendar	0.11	0.27	Phone	0.07	0.08
Chair 1	0.02	0.30	Poster Mystrands	0.71	0.12
Chair 2	0.01	0.34	Poster spices	0.05	0.05
Chair 3	0.02	0.05	Rack	0.06	0.55
Charger	0.0	0.08	Red cup	0.01	0.05
Cube 1	0.06	0.21	Stapler	0.02	0.20
Cube 2	0.0	0.56	Umbrella	0.05	0.58
Cube 3	0.03	0.24	Window	0.10	0.08
Extinguisher	0.09	0.13	Wine bottle	0.03	0.32
Monitor 1	0.02	0.01			

TABLE II
Recall and precision values for each object category for the Viola and Jones object detector when using a training set of several good quality images per object and with synthetically generated images.

The dataset used for the first test only had a few images for each type of object: 50 to 70 images per class. In Table I the results obtained for sequences IIIA30-1 and IIIA30-2 are shown. With so few training data, the Viola and Jones classifier is able to find only some instances for objects of 11 out of the 29 categories. This performance is expected due to the limited amount of training data.

Table II shows the results obtained with twenty good quality training images, but further enhancing the set by synthetically generating a hundred extra images for each training sample. As it can be seen, the usage of high quality images and the synthetic views significantly improved the results.

Finally, Table III shows the results obtained using the third training set, which consisted of hundreds of good quality im-

	All		Non-Occluded		Occluded	
Object	Recall	Prec	Recall	Prec	Recall	Prec
Grey battery	0.36	0.24	0.41	0.24	0.0	0.0
Red battery	0.37	0.82	0.44	0.82	0.0	0.0
Bicycle	0.0	0.0	0.0	0.0	0.0	0.0
Ponce book	0.81	0.88	0.86	0.86	0.25	0.02
Hartley book	0.66	0.94	0.70	0.94	0.0	0.0
Calendar*	0.33	0.08	0.38	0.08	0.0	0.0
Chair 1	0.0	0.0	0.0	0.0	0.0	0.0
Chair 2*	0.0	0.0	0.0	0.0	0.0	0.0
Chair 3	0.0	0.0	0.0	0.0	0.0	0.0
Charger	0.12	0.08	0.12	0.08	0.0	0.0
Cube 1	0.22	0.43	0.23	0.29	0.2	0.15
Cube 2	0.23	0.11	0.20	0.09	0.34	0.03
Cube 3	0.28	0.53	0.37	0.48	0.09	0.06
Extinguisher	0.0	0.0	0.0	0.0	0.0	0.0
Monitor 1*	0.0	0.0	0.0	0.0	0.0	0.0
Monitor 2*	0.23	0.57	0.39	0.57	0.0	0.0
Monitor 3*	0.04	0.13	0.05	0.13	0.0	0.0
Orbit box*	0.15	0.03	0.17	0.03	0.0	0.0
Dentifrice	0.0	0.0	0.0	0.0	0.0	0.0
Poster CMPI	0.11	0.34	0.19	0.34	0.0	0.0
Phone	0.05	0.09	0.0	0.0	0.3	0.09
Poster Mystrands	0.0	0.0	0.0	0.0	0.0	0.0
Poster spices	0.04	0.38	0.12	0.38	0.0	0.0
Rack	0.0	0.0	0.0	0.0	0.0	0.0
Red cup	0.89	0.89	0.89	0.89	0.0	0.0
Stapler	0.24	0.21	0.24	0.21	0.0	0.0
Umbrella	0.0	0.0	0.0	0.0	0.0	0.0
Window	0.03	0.40	0.10	0.40	0.0	0.0
Wine bottle*	0.10	0.06	0.10	0.06	0.0	0.0

TABLE III
RECALL AND PRECISION VALUES FOR EACH OBJECT CATEGORY USING THE VIOLA & JONES OBJECT DETECTOR AND THE THIRD TRAINING SET DESCRIBED . WHEN WE DECOMPOSE THE PRECISION-RECALL VALUES FOR OCCLUDED AND NON-OCCLUDED OBJECTS, RESULTS SHOWS A PERFORMANCE DROP FOR OCCLUDED OBJECTS. THE ASTERISK MARK DENOTES OBJECTS TRAINED FROM SYNTHETIC IMAGES.

ages extracted from video recordings done with a conventional camera.

A conclusion that can be quickly inferred from the table is the decrease in performance caused by occlusions. Even objects that achieve a good recall and precision with good viewing conditions, fail in the case of occlusions. In contrast, blurring and illumination variations did not affect performance significantly. Regarding the object types, (textured, untextured and repetitively textured) textured objects obtained an overall recall of 26% and precision of 33%, similar to that of repetitively textured objects (24% recall and 36% precision). Finally, untextured objects obtained 14% of recall and 19% precision.

The performance on the posters is surprisingly low, as they are usually considered "easy" objects. The most probable explanation is the large changes in point of view that the posters suffer through the video sequences. The time necessary to apply the classifiers for all the classes to one test image is 728 ms on average.

IV. CONCLUSIONS

We have presented a publicly available and hard object detection dataset acquired with a mobile robot, that faithfully represents the typical problems encountered in mobile robotics and mobile computing in general (i.e. low resolution, motion blur, etc.). The dataset contains three sequences of varying length, with bounding box annotations for 29 challenging object types, as well as good quality training images.

In order to set a baseline for future evaluations, we have run the Viola and Jones Cascade of classifiers object detector on the three sequences. This study is part of a larger work evaluating three state of the art object detectors suitable for mobile robotics [10]: the SIFT object recognition system [11], the Vocabulary Tree method [12] and the Viola and Jones Cascade of Classifiers method [5].

Despite the use of very simple image features, the Viola and Jones Cascade of classifiers attains a good level of recall for several objects in a very low runtime. Its main drawbacks are the large (in comparison with other techniques) training dataset required to obtain a good level of performance, and the limited robustness to changes in the point of view and occlusions of the method, as well as a significant number of false positives that have to be filtered out in later stages. Furthermore, some theoretically "easy" objects, such as the posters, proved to be troublesome to the Viola and Jones method. This is probably due to overfitting to some particular view, or to too much variability of the very rich Haar feature distribution when changing the point of view, where the method was unable to find any recognizable regular pattern.

Nevertheless, the idea of a boosted cascade of weak classifiers is not limited to the very fast but simple Haar features, but any kind of classifier can be used for that matter. A very interesting alternative is using linear SVMs as weak classifiers, since it allows to add a non-linear layer to an already efficient linear classifier. Such idea has been already successfully applied in a few cases [13], [14], and we believe it is a very interesting line to investigate.

ACKNOWLEDGMENTS

This work was supported by the FI grant and by the 2009-SGR-1434 project from the Generalitat de Catalunya, by the Spanish Ministry of Education and Science under projects TRA2010-21371-C03-01 and Consolider Ingenio 2010: MIPRCV (CSD200700018) and by the Rio Tinto Centre for Mine Automation and the ARC Centre of Excellence programme, funded by the Australian Research Council (ARC) and the New South Wales State Government. .

REFERENCES

[1] S. Vasudevan, S. Gachter, V. Nguyen, and R. Siegwart, "Cognitive maps for mobile robots - an object based approach," in Robotics and Autonomous Systems, Volume 55, Issue 5, From Sensors to Human Spatial Concepts, 31 May 2007, Pages 359-371., 2007.

[2] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[4] N. Pinto, D. D. Cox, and J. J. Dicarlo, "Why is real-world visual object recognition hard?" *PLoS Computational Biology*, vol. 4, no. 1, pp. e27+, January 2008. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.0040027

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, p. 511.

[6] R. Lienhart, E. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," in *In DAGM 25th Pattern Recognition Symposium*, 2003, pp. 297–304.

[7] A. Torralba, K. Murphy, and W. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 854–869, 2007.

[8] M. Jones and P. Viola, "Fast multi-view face detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. Citeseer, 2003.

[9] C. Huang, H. Ai, B. Wu, and S. Lao, "Boosting nested cascade detector for multi-view face detection," in *Proceedings of the 17th International Conference on Pattern Recognition*, 2004, pp. 415–418.

[10] A. Ramisa, "Localization and object recognition for mobile robots," Ph.D. dissertation, Universitat Autonoma de Barcelna, 2009.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Interantional Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," *Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 2161–2168, 2006.

[13] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *International Journal of Computer Vision*, vol. 63. Springer, 2005, p. 153161. [Online]. Available: http://www.springerlink.com/index/T61K38U53J531344.pdf

[14] D. Aldavert, A. Ramisa, R. Toledo, and R. L. D. Mantaras, "Fast and Robust Object Segmentation with the Integral Linear Classifier," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.