Two-Stage Part-Based Pedestrian Detection

Møgelmose, Andreas; Prioletti, Antonio; Trivedi, Mohan M.; Broggi, Alberto; Moeslund, Thomas B.

Published in:
15th International Conference on Intelligent Transportation Systems

DOI (link to publication from Publisher):
10.1109/ITSC.2012.6338898

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

2012 15th International IEEE Conference on Intelligent Transportation Systems
Anchorage, Alaska, USA, September 16-19, 2012

Two-stage Part-Based Pedestrian Detection

Andreas Møgelmose, Antonio Prioletti, Mohan M. Trivedi, Alberto Broggi, and Thomas B. Moeslund

Abstract— **This paper introduces a part-based two-stage pedestrian detector. The system finds pedestrian candidates with an AdaBoost cascade on Haar-like features. It then verifies each candidate using a part-based HOG-SVM doing first a regression and then a classification based on the estimated function output from the regression. It uses the Histogram of Oriented Gradients (HOG) computed on both the full, upper and lower body of the candidates, and uses these in the final verification. The system has been trained and tested on the INRIA dataset and performs better than similar previous work, which uses full-body verification.**

I. Introduction

Pedestrian detection is currently a very large research field. It can be used in surveillance, Advanced Driver Assistance Systems (ADAS), and many other places. The ADAS scenario offers plenty of challenges (as summarized in [1]): High variability in appearance among pedestrians, cluttered backgrounds, highly dynamic scenes with both pedestrian and camera motion, and strict requirements in both speed and reliability. Input from a reliable pedestrian detection system can be used to warn the driver about people in front of the car (a warning that must not overload the driver with information [2]), prepare or even activate a braking maneuver to prevent a collision, or deploy other safety systems such as airbags.

ADAS is a a challenging domain to work within. Braking systems take a short while to apply, and reaction times must be fast for driving, where fractions of a second can be the deciding factor between a collision and a near-miss. At the same time, the system must be robust, so the braking system is not deployed mistakenly (due to a false positive detection), which could itself lead to accidents, or worse, not deployed at all (due to a missed detection). Further reasoning than just detection is necessary in such a framework, with pedestrian intent estimation being a good example, as presented in [3] and reviewed in [4] or as another example, automatic breaking as in [5].

The approach presented in reference [6] is a combination of a Haar based boosted cascade classifier for high speed with a HOG-SVM detector for reducing false positives. The approach that we describe in this paper, can be viewed as an extension of this idea. In our approach, we extend such a combination idea with to a part-based solution, which lowers the false positive rate even further. The part-based philosophy has never before been applied to this detection scheme.

This paper is structured as follows: In the next section, we describe some of the work related to ours and we provide an overview of our algorithm. In the next sections we describe each stage in the algorithm in detail. Finally, in V, we describe the performance of our algorithm followed by suggestions for future work and a conclusion.

II. General approach and related work

As mentioned, pedestrian detection is a field with much attention from the research community. Even when narrowed to applications in connection with cars and ADAS, a large body of work exists. A classic method of pedestrian detection is a boosted cascade on Haar-like features, first presented by Viola and Jones [7]. It is very fast, but lacks robustness due to the high appearance variability among pedestrians in the real world. Instead, many people turn to the HOG-SVM solution presented by Dalal and Triggs [8]. It is much more robust and generally detect pedestrians in harder situations, while keeping a low number of false positives. Its problem lies in processing speed. As mentioned, the ADAS application requires fast processing, something that is not immediately obtainable with the HOG-SVM detector. The HOG-SVM method was explored for use with infrared images in [9]. For further exploration of pedestrian detectors, we refer the reader to the general survey by Gerónimo et. al. [1] or, for vision-only based systems, Gandhi and Trivedi [10], [11], and Krotosky and Trivedi [12]. The system presented in our paper uses monocular vision as base for the detection. This means that the hardware requirements for the car are low and realistically possible - many cars are already outfitted with a front facing camera for other purposes, such as lane detection. For a survey of monocular vision based methods, see [13].

We combine the speed of the Haar detector with the robustness of a part-based HOG-SVM detector. The base for the method used in this paper was first presented by Geismann and Schneider [6], but is also covered by others in various versions [14], [15]. Apart from using a combination of a Haar-cascade and HOG-SVM, Geismann and Schneider also evaluated using a sparse HOG descriptor to speed up the verification. Part-based pedestrian detection has been presented in other contexts before, such as [16], [17], [18].

The properties of the Haar cascade and the HOG-SVM detector make them prime candidates for combination: The Haar cascade does the initial pass, finding Regions Of

A. Møgelmose is a research scholar at the CVRR Lab, UCSD and PhD student at the VAP Lab, AAU, Denmark. am@create.aau.dk

A. Prioletti is a research scholar at the CVRR Lab, UCSD and master of science in computer vision from Vislab, University of Parma, Italy. antonio.prioletti@studenti.unipr.it

M. M. Trivedi is with the CVRR Lab at University of California, San Diego mtrivedi@ucsd.edu

A. Broggi is with Vislab at University of Parma, Italy broggi@vislab.it

T. B. Moeslund is with the VAP Lab at Aalborg University, Denmark tbm@create.aau.dk

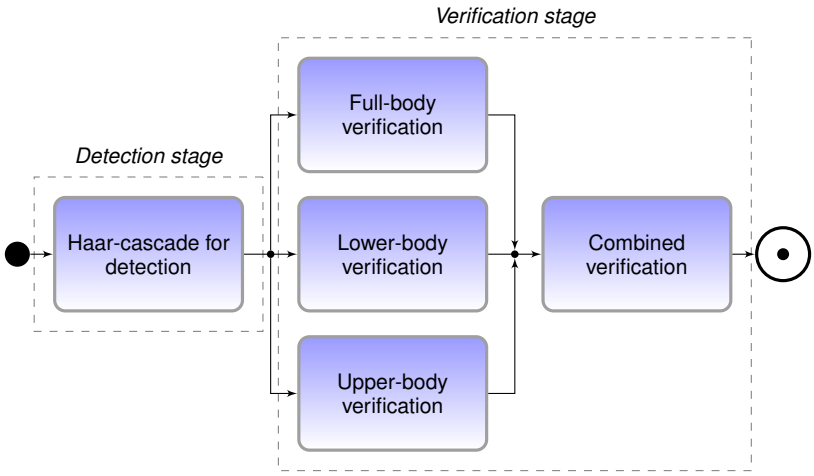


Fig. 1. The flow of the algorithm described in this paper.

Interest (ROIs) that are passed on to the HOG-SVM detector which verifies the initial findings by the Haar cascade. The first stage is called the *detection stage* and the second the *verification stage*. That is the basics of the approach outlined in [6]. As mentioned earlier, the Haar cascade is not very robust, but that is not a problem, since we use it only for determination of ROIs, so we can allow many false positives, which also drives up the number of true positives to an acceptable level.

Our goal is to lower the number of false positives without too much penalty in the detection rate. In order to do this, we alter the verification stage to not only verify based on a full body classification, but also a lower body and upper body classifier. We combine these results to figure out whether the ROI contains a person.

The combination of verification results is done in two ways which are compared: A simple majority vote, requiring at least two of three classifiers to verify the detection, and a more advanced way which introduces a third stage to the algorithm, classifying each window based on the estimated function value from an SVM regression performed on each part.

An overview of the flow through the algorithm can be seen in fig. 1.

III. Detection stage

The detection stage is an AdaBoost cascade on Haar-features [7]. It works by using AdaBoost to learn a number of weak classifiers, which are combined into strong classifiers. Several layers (called stages) of these strong classifiers are then combined in a cascade to create the final detection. The cascaded structure makes the algorithm very fast, since most candidates are discarded in one of the first stages, thus not having to be calculated in following stages. Only the actual detections have to pass through all stages. The algorithm is described in detail in [7].

Throughout this paper, we work with the INRIA Pedestrian Dataset [8]. Thus, the detection cascade was trained with the training set given therein: 2416 positive images and

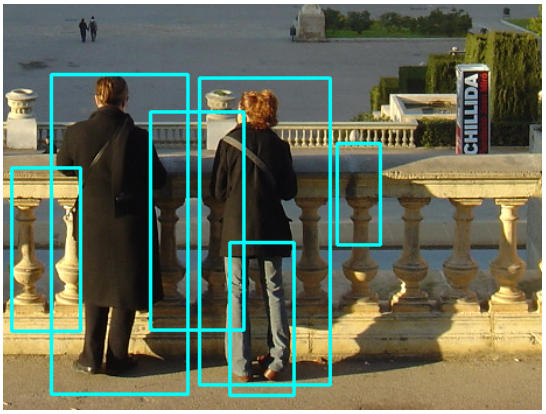


Fig. 2. Example of the output from the detection stage. It is clear that it contains several false positives, but that is desired, since it ensures that also the true positives are included.

12180 negative images. The training images were cropped closely around the annotated persons, because Haar-cascades does not benefit from having as much background included as HOG-based classifiers. After the crop the training images were resized to 12x28 pixels.

The detection stage is set up so that it finds the maximum possible number of pedestrians, which also means that it will return plenty of false positives. A larger number of false positives will slow down the computation, since the verification stage must process more, but it is a worthy trade-off given that the true positive rate of this stage forms the upper bound of detections for the entire system.

The detection stage returns bounding boxes of all the potential pedestrians in the picture, which are sent on to the verification stage. Part based detection in the detection stage is not used, since the data from [19] shows that the Haar cascade generally performs bad in part-based detection schemes. An example of the output of the detection stage can be seen on fig. 2.

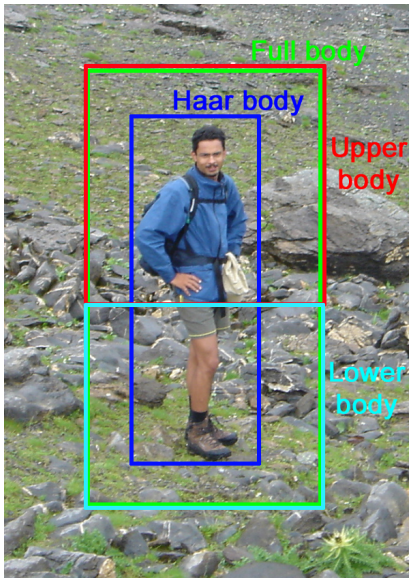


Fig. 3. The four types of training images used in this system: The three parts for the verification stage, and a closer crop for the detection stage.

IV. Verification stage

The part-based verification stage used in this work differs from the full-body verification stage of Geismann and Schneider's [6]. We use a part-based detection scheme. The verification stage consists of two sub-stages: The individual part verification and the combined verification. Three SVM regressions based on dense HOG descriptors are calculated and applied to the ROIs given by the detection stage. One is for full body classification, one is for lower body classification, and one is for upper body classifications.

Our algorithm uses classic dense HOG descriptors (as opposed to the sparse descriptors used in [6]). They are calculated using integral images in an effort to speed up the process, as described in [20]. Since HOG works best if some amount of background is introduced to the detection window, the ROIs are resized appropriately from the tight boxes that are returned by the detection stage. Then the content of the ROIs is scaled so it matches the size the SVMs were trained with. At this point the HOG is calculated and passed on to the SVMs.

As in the detection stage, each SVM is trained with the INRIA training set. The full body SVM was trained with the full training images, whereas the lower- and upper-body SVMs were trained with the lower and upper half of the training images, respectively. In our system, there is no overlap between the lower and upper body. The parts of training images used for each type are shown in fig. 3. So in total, three SVMs were used.

To do the combined verification, two different methods were tested: Majority voting and regression output classification.

For majority voting, a regular SVM for classification was trained. It returns which class (pedestrian vs. non-pedestrian) the current detection window belongs to. If at least two out of three classifiers label the window as a pedestrian, it is described as a detection. If a detection is labeled 1 and no detection is labeled -1, the formula used for the majority voting is:

$$l_{out} = \begin{cases} 1 & \text{if} \sum_{i=0}^{i<3} l_i \geq 1 \\ -1 & \text{if} \sum_{i=0}^{i<3} l_i < 1 \end{cases} \quad (1)$$

where l_{out} is the final decision and l_i is the output from one of the three part-based detectors.

For regression output classification, the three part SVMs were instead trained for regression. The training was performed so the resulting function would ideally return 1 in the case of a detection and -1 when nothing was found. When an unknown window is passed through the output function, it will return a value close to 1 if it is a pedestrian, and a value close to -1 otherwise. The output of these three regressions create their own 3 dimensional feature space. Another SVM has been trained to classify in this space. The output from the three regressions is passed into this second SVM and the output from that classifier is the final label.

TABLE I

Overview of the detection rates achieved by Geismann and Schneider [6] with 0.2 false positive per frame

Video	1	2	3	4	5	Mean
Dense descriptor	52%	70%	91%	61%	55%	65.8%
Sparse descriptor	45%	53%	85%	69%	58%	62%

V. Experiments and test

In order to set various parameters so that the best possible performance is achieved, several experiments have been performed. While the training part of the INRIA dataset was used to train both the detection stage and the verification stage, the test part has been used as base for these experiments. It contains 742 images in total, of which 289 contain one or more persons. In total the test set contains 589 persons that should be detected by a perfect system.

The baseline for the comparison is the performance of the system in a configuration similar to the one by Geismann and Schneider: Two stages, but no part-based verification. The principal results from their paper can be seen in table I. Each of the five results in the table are from a test video they obtained from a driving car. Unfortunately we do not have access to the test videos they used, so our results cannot be compared directly with those. Instead we compare the performance of our own implementation of their algorithm to out part-based algorithm.

One of the most important parameters in the system is the number of stages in the Haar-cascade, in this paper designated k. It is interesting to see what impact the changes in k has on the complete system. In fig. 4, an ROC curve is shown for the full system with varying depths in the detection stage. The importance of this parameter is evident. As k is lowered, the number of detections rise, but at a large cost in false positives. The final system uses $k = 15$, since it seems

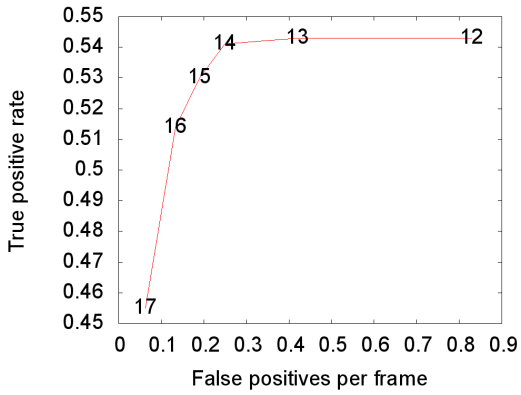


Fig. 4. Receiver-Operating-Characteristic for the full system with varying k, cascade depths in the detection stage.

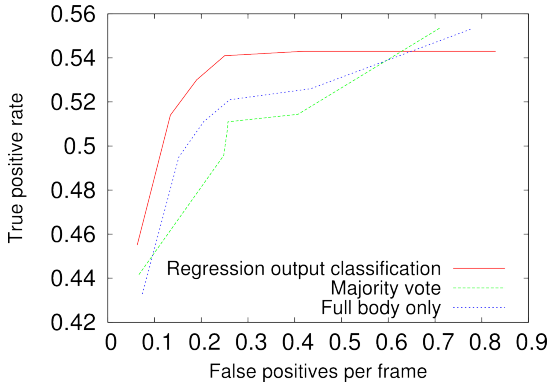


Fig. 5. Receiver-Operating-Characteristic for the final system. The majority voting approach is not performing better than the full-body approach, but the regression classification approach is consistently better.

to give an acceptable trade-off between true positives and false positives.

The choice of k has an impact on the speed of the system, since more detection windows means slower performance. The full (but non-optimized) system has been run with several numbers of stages and timed, to get a sense for the speed effects it might have. The results are seen in table II. While the detection speed here are not overwhelmingly fast, it is worth to note that the test images are of a relatively high resolution and a production system could easily be run with smaller images.

Another important parameter is the padding, p: The amount with which the ROIs returned by the detection stage is enlarged with. The HOG-SVM detector works better if more background is included than what the Haar-cascade uses, so there is no question that the ROIs must be enlarged. Experiments showed that a padding of 3 performed best. When the padding rises, the pedestrian in the ROI becomes a lot smaller relative to the ROI, than the pedestrians in the training set. Because HOG-SVM is not scale invariant, that will alter the output of the detector and some testing was required to make it work properly. The padding value itself is used to calculate the padding in pixels to apply to the ROI. The width in pixels is calculated as:

$$p_{pixels} = \frac{w_{ROI}}{w_t} \cdot p \tag{2}$$

where p is the padding value, w_{ROI} is the width of the found ROI, w_t is the width of the training images, and p_{pixels} is the padding measured in pixels. The padding is applied on all four sides of the image.

After introducing part-based verification to the system, experiments were made to determine whether the simple majority voting or the confidence classification worked the best. These tests were done with the best settings, as determined earlier in this section. The results are shown in fig. 5. In absolute numbers, the detection rate is decent, though not spectacular. The important part is the difference between the old two-stage approach with only full-body verification and the new approach. While the voting based approach is not any better then the old full-body verification, the part-based version with regression output classification is better all across the range of false positives per frame.

Examples of detections can be seen in fig. 6.

TABLE II

CHANGES IN PROCESSING SPEED FOR DIFFERENT VALUES OF k

Number of images	100
Mean image width	711.88 pixel
Mean image height	818.72 pixel

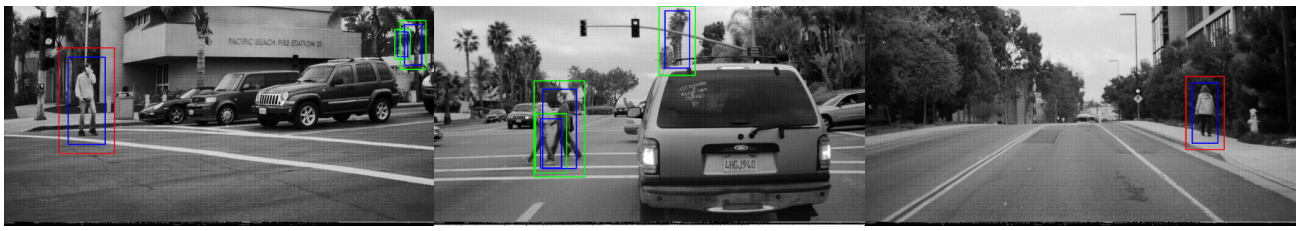
k	**Mean time per frame**
12	2.5 s
13	1.84 s
14	1.57 s
15	1.46 s
16	1.39 s
17	1.22 s

VI. FUTURE WORK

The algorithm has a series of parameters that can be adjusted to enhance performance. In this work, a few tests and comparisons has been carried out, in order to give the best performance. However, a more formal investigation of the optimal parameters would be interesting. One possibility is to use a genetic algorithm or particle swarm optimization to set the best parameters.

This work has mostly been concerned with lowering the number of false positives in the classic combination of a Haar cascade and HOG-SVM, so speed has not been a primary concern. First and foremost, several optimizations, such as using sparse HOG calculation, are presented in [6], and they could be implemented with little impact on performance.

The system presented here deals only with single-frame detection. A full pedestrian detection system would very likely benefit from using tracking between frames to enhance the performance.



(a) (b) (c)

Fig. 6. Example outputs of the detector. Input images are images captured using one of LISA's experimental cars. Red boxes indicate a detection, blue are the candidates from the detection stage, and green are the candidates with added padding. In (a) the pedestrian is detected, while a couple of false candidates from the detector are ignored. In (b) the pedestrians are not detected, since the detection stage does not find accurate enough candidate boxes. In (c) the pedestrian is detected and no false windows are found in the detection stage.

VII. Concluding remarks

In this paper, a part-based two-stage pedestrian detector has been presented. It builds on previous work by Geismann and Schneider [6], but extends it by introducing a part-based verification system instead of just a full body verification. The system works in two stages: A detection stage based on an AdaBoost cascade on Haar-like features. Its purpose is to find all pedestrian candidate patches in the input image. All these Regions Of Interest are sent on to a verification stage, where the Histogram Of Oriented Gradients (HOG) is computed for the entire person, the lower body, and the upper body. Each of the HOGs are then sent trough an SVM that computes a confidence value for which class (pedestrian or non-pedestrian) the part belongs to. These values are then passed into a second SVM-classifier, which performs the final verification. The system has been tested on the INRIA dataset and the results show that when compared with the original two-stage detector, it performs better across the full range of false positives per frame.

VIII. Acknowledgment

The authors would like to thank our colleagues in the LISA-CVRR lab and Vislab for feedback during the work. Especially valuable were the comments from Dr. Brendan Morris and Sayanan Sivaraman.

References

[1] D. Geronimo, A. Lopez, A. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 7, pp. 1239–1258, 2010.

[2] B. Morris and M. Trivedi, "Vehicle Iconic Surround Observer: Visualization platform for intelligent driver support applications," in *Intelligent Vehicles Symposium (IV), 2010 IEEE*. IEEE, 2010, pp. 168–173.

[3] T. Gandhi and M. Trivedi, "Image based estimation of pedestrian orientation for improving path prediction," in *Intelligent Vehicles Symposium, 2008 IEEE*. IEEE, 2008, pp. 506–511.

[4] A. Doshi and M. Trivedi, "Tactical Driver Behavior Prediction and Intent Inference: A Review," in *14th IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2011.

[5] A. Broggi, P. Cerri, S. Ghidoni, P. Grisleri, and H. Jung, "A new approach to urban pedestrian detection for automatic braking," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 10, no. 4, pp. 594–605, 2009.

[6] P. Geismann and G. Schneider, "A two-staged approach to vision-based pedestrian recognition using Haar and HOG features," in *Intelligent Vehicles Symposium, 2008 IEEE*. IEEE, 2008, pp. 554–559.

[7] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2001.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. Ieee, 2005, pp. 886–893.

[9] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," in *Intelligent Vehicles Symposium, 2006 IEEE*. Ieee, 2006, pp. 206–212.

[10] T. Gandhi and M. Trivedi, "Pedestrian collision avoidance systems: A survey of computer vision based recent studies," in *Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE*. IEEE, 2006, pp. 976–981.

[11] ——, "Pedestrian protection systems: Issues, survey, and challenges," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, no. 3, pp. 413–430, 2007.

[12] S. Krotosky and M. Trivedi, "On Color-, Infrared-, and Multimodal-Stereo Approaches to Pedestrian Detection," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, no. 4, pp. 619–629, dec. 2007.

[13] M. Enzweiler and D. Gavrila, "Monocular pedestrian detection: Survey and experiments," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 12, pp. 2179–2195, 2009.

[14] X. Yuan, X. Shan, and L. Su, "A Combined Pedestrian Detection Method Based on Haar-Like Features and HOG Features," in *Intelligent Systems and Applications (ISA), 2011 3rd International Workshop on*. IEEE, 2011, pp. 1–4.

[15] W. Yongzhi, X. Jianping, L. Xiling, and Z. Jun, "Pedestrian Detection Using Coarse-to-Fine Method with Haar-Like and Shapelet Features," in *Multimedia Technology (ICMT), 2010 International Conference on*. IEEE, 2010, pp. 1–4.

[16] X. Mao, F. Qi, and W. Zhu, "Multiple-part based Pedestrian Detection using Interfering Object Detection," in *Natural Computation, 2007. ICNC 2007. Third International Conference on*, vol. 2. IEEE, 2007, pp. 165–169.

[17] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 90–97.

[18] ——, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, 2007.

[19] I. Alonso, D. Llorca, M. Sotelo, L. Bergasa, P. de Toro, J. Nuevo, M. Ocaña, and M. Garrido, "Combination of feature extraction methods for SVM pedestrian detection," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, no. 2, pp. 292–307, 2007.

[20] F. Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 829–836.