Pedestrian Detection Optimization Based on Random Filtering

Victor Hugo Cunha de Melo, Samir Leão, William Robson Schwartz
Universidade Federal de Minas Gerais
Department of Computer Science
Belo Horizonte, Minas Gerais, Brazil
Email: {victorhcmelo, samirleao, william}@dcc.ufmg.br

***Abstract*—The large number of surveillance cameras available nowadays in strategic points of major cities provides a safe environment. However, the huge amount of data provided by the cameras prevents the manual processing requiring the application of automated methods. Among such methods, pedestrian detection plays an important role in reducing the amount of data by locating only the regions of interest for further processing regarding activities being performed by agents in the scene. However, pedestrian detection methods currently available are unable to process such large amount of data in real time. Therefore, optimization techniques have to be employed to allow real-time detection even when large volumes of data need to be processed. Focusing on optimization, this work proposes a novel approach that performs a random filtering in the image to discard a large number of detection windows quickly allowing a reduction in the computational cost.**

***Keywords*-pedestrian detection optimization; random filtering; visual surveillance;**

I. Introduction

Pedestrian detection plays a major role in applications such as surveillance due to the need of processing a large amount of data captured from multiple cameras and locating the agents that are performing some activity of interest. Therefore, the volume of data capture can be reduced focusing only on those regions of interest so that problems such as person tracking, face recognition, person re-identification, action and activity recognition can be solved and the activities performed by the agents in the scene might be analyzed.

Several challenges are faced by the pedestrian detection problem [1]. Among them are the changes in appearance due to different types of clothing, illumination changes and pose variations, the low quality of the data acquired, and the possible small size of the pedestrian on it, which makes the detection process harder. In addition, a large number of applications require a high performance and reliable detection results, which increases the need for efficient and accurate pedestrian detection approaches.

Even though many pedestrian detection have been proposed, most of the times they need to be employed on domains in which millions of images need to be processed quickly. However, the methods currently available are not able to provide such performance [2]. Therefore, the development of methods to reduce the computational cost significantly is desirable. One way of achieving that is to focus on optimization approaches. The most common are those based on cascade of rejection, region of interest filtering and GPU-based processing to reduce the number of detection windows to be evaluated.

This work proposes a novel approach to optimize the detection by performing a random filtering in the image to discard a large number of detection windows and therefore reduce the computational cost. The experimental evaluation shows in two publicly available datasets (ETHZ and INRIA) that it is possible to discard a large number of detection windows and still achieve very accurate results. In addition, future directions for this work, based on learning the object distribution for performing further rejection of windows in the scene, are remarked.

II. Related Work

Several methods address the aforementioned problem induced by dense search in sliding window approaches. Some proposed heuristics evaluate windows in fixed sizes, which are subsampled for certain strides [3], [4]. The larger the stride, the more sparse the sampling of the image will be.

A common approach used to optimize object detection is based on a cascade of rejection, composed by weak classifiers learned during the training. The main idea behind this approach is to use simple classifiers to discard detection windows that are easy to classify. While the remaining windows of a stage advance through the cascade, more complex classifiers are used. Viola and Jones [5] propose an object detector using this technique by successively combining classifiers with increasing complexity, building the cascade stages and rejecting a large amount of windows in early stages. Applying a similar approach Zhu *et al.* [6] performed the extraction of HOG descriptors using a real time processing framework and achieved a detection rate comparable to that achieved by Dalal and Triggs [3].

Focusing on searching only promising regions of the image, Lampert *et al.* [7] propose a method to perform object localization relying on a branch-and-bound approach that finds the global optimum of a quality function over every possible subimage. It returns the same object locations that an exhaustive sliding window approach would. At the same time it requires fewer classifier evaluations than there are candidate regions in the image, typically running in linear time or faster.

Saliency detectors are able to detect regions of interest by simulating the behavior of the human visual system. In the first phase, called pre-attentive visual search, they quickly detect the possible positions of proto-objects in the image. The obtained saliency map suggests the position of the proto-objects. Feng *et al.* [8] propose a filtering method that finds the salience of each window and segments the image into regions based on their similarities. In order to find the most probable window to contain a salient object, it is employed the difference among regions given their LAB histograms and spatial distances.

For detection of objects in different sizes, [9], [10] propose the direct analysis of features extracted in multiple scales of the image. Based on saliency detectors, Grimaldo *et al.* [11] propose a method based in multi-scale Spectral Residual Analysis (MSR), in which an image is resized by several times a factor to cover different scales. In each resizing, a saliency map is created and a sliding window approach is applied, then a quality function is computed in each map in order to discard regions. In comparison with a regular sliding window approach, the MSR method was able to reduce in 75% the number of windows to be evaluated by an object detector and improving the detection rate in most cases.

For accelerated computation and improvement of object detectors, the multicore architecture of GPUs can be very handy. In order to do that, the algorithms must be designed to explore those hardware features. Pedestrian detection on GPUs can improve the feature extraction, since it is a stage of high computational cost. Wojek *et al.* [12], Zhang *et al.* [13] and Masaki *et al.* [14] demonstrated efficient parallel techniques using GPU in order to do HOG features extraction.

III. Proposed Approach

The first step of the proposed method consists of sampling a set of detection windows in several scales using a sliding window approach. The sampling procedure starts from a minimum width and height, generating several windows of fixed size from the image. The entire image is scanned by this detection window size, which is moved by a stride x and y—computed as a percentage of the height and width of the window, respectively. After the entire image has been scanned, the detection window size is resized by a scaling factor and the procedure is repeated until the detection window's size reaches a maximum width and height.

With the set of all sampled detection windows, the next step is to discard a random subset of these windows. In this work, we consider that the pedestrians are uniformly distributed in the image, hence we are performing an uniform randomly removal of windows. However, it is possible to see in Section IV that the pedestrians' distribution does not follow a uniform distribution. As future work, we intend to exploit that knowledge to discard windows based on a learned distribution, maintaining more windows in regions more likely to contain pedestrians and discarding windows in regions with low probability of pedestrian presence. After discarding the detection windows, the next step is to feed the detector with the selected windows.

Why does the random rejection of detection windows work? The problem of classifying windows as containing pedestrians or not may be seen as a task of finding an optimal subset of windows containing humans from a finite set of windows. As most maximum search problems, the exact solution is computationally expensive. Instead, it is possible to find almost optimal approximate solutions by using probabilistic methods as the one described as follows.

The problem at hand might be formulated as: given a set of m windows, where $M := \{f_1, \ldots, f_m\}$, and $\mathcal{Q}[f]$ a criterion to evaluate whether a detection window is covering a portion with a pedestrian, the problem consists in finding a window $\hat{f}$ that maximizes $\mathcal{Q}[f]$. In pedestrian detection, one is interested in finding not only the window $\hat{f}$ that maximizes $\mathcal{Q}[f]$, but also a subset of windows with the largest $\hat{f}$, since more than one pedestrian might be in the image.

To solve the aforementioned problem, all terms $\mathcal{Q}[f_i]$ must be computed, which demands m detection window evaluations. Due to the multiple scales that are considered to locate all pedestrians in the scene, the number of extracted windows is large for a given image, rendering this operation too expensive. For instance, for an image with dimensions 640×480 pixels, there are approximately 60,000 detection windows that need to be evaluated to detect pedestrian in multiple scales. Therefore, it is imperative to find a cheaper approximate solution.

Schölkopf *et al.* [15] demonstrated that, selecting a random subset $\tilde{M} \subset M$ that is sufficiently large, one can take the maximum over $\tilde{M}$ as an approximation of the maximum over M. If a small fraction of $\mathcal{Q}[f_i]$ whose values are significantly smaller or larger than the average does not exist, one can obtain a solution that is close to the optimum with high probability.

To compute the required size, $\tilde{m} = |\tilde{M}|$ $(\tilde{M} \subset M)$, of a random subset to achieve a desired degree of approximation, Schölkopf *et al.* show that one can use the following equation

$$\tilde{m} = \frac{\log(1-\eta)}{\ln(n/m)} \tag{1}$$

where η is the desired confidence and n denotes the number of elements in M having $\mathcal{Q}[f]$ smaller than the maximum of $\mathcal{Q}[f]$ among the elements in $\tilde{M}$.

In the pedestrian detection problem based on sliding windows, one human is covered by more than one detection window leading to a correct detection. That behavior is due to the redundancy resulting from the small strides in x and y and multiple scales. In fact, the number of correct windows increases linearly with the number of detection windows in the image. For instance, given an image with $m = 60{,}000$ detection windows uniformly sampled in an image 640×480, 583 windows will contain the correct location of a pedestrian (high $\mathcal{Q}[f_i]$). According to the theorem, a random sample with $\hat{m} = 133$ will have a 95% probability that at least one of 583 windows contains pedestrians. Figure 1 illustrates that, by

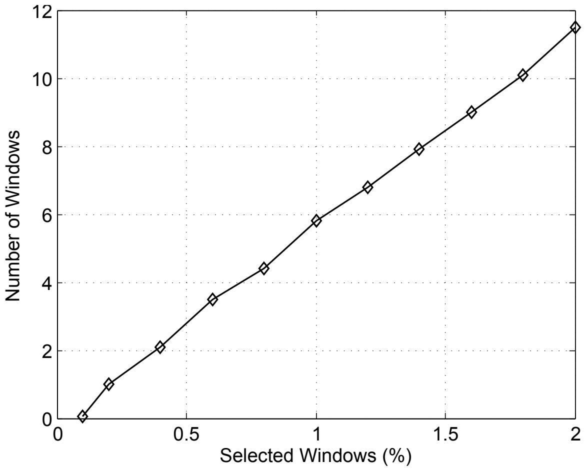


Fig. 1: Mean number of detection windows covering a pedestrian in the INRIA dataset. As standard approach to determine whether a pedestrian is covered by a detection window, we consider that a window A covers a ground-truth window B when their intersection divided by their union is greater than 0.5.

showing the number of windows covering a pedestrian when a small percentage of the detection windows is selected at random.

Previously, Zhu *et al.* [6] have adopted this theorem during their training phase of a cascade of rejection. Each stage would require to evaluate $5{,}301$ blocks to be considered, which is very time consuming. To reduce the number of evaluated blocks, they applied this sampling theorem and selected only 250 blocks at random each round.

IV. Experimental Evaluation

This section analyzes the impact of the proposed methodology. The experimental evaluation focus on three main aspects.

The first aspect examines if there is no pedestrian missing after applying uniform random filtering. This can be determined by applying random filtering and evaluating the results obtained with respect to the ground-truth.

Even though the method may obtain high recall rates regarding the ground-truth, the windows are not necessarily centralized over the pedestrians. They might be displaced by an offset. To ensure that a pedestrian detection method may detect pedestrians in these dislocated windows, it is analyzed the trade off between the amount of subwindows randomly discarded and the results of a pedestrian detector over these windows.

Finally, the last aspect explores the distribution of the pedestrians' centroids into the datasets. This is a preliminary study for future works.

Experimental setup. The windows are sampled using a minimum width and height of 28×60. The window size is rescaled by a factor of 1.15 until it reaches the maximum width and height of 260×700. The windows are displaced by a stride of 12% and 4% of the width and height, respectively. Two different datasets were used in this evaluation, the ETHZ pedestrian dataset [16] and INRIA Person Dataset [3].

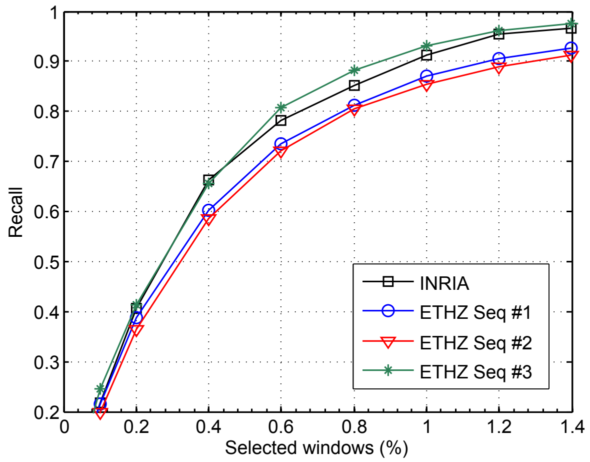


Fig. 2: Achievable recall as a function of variable number of selected windows, evaluated on INRIA and ETHZ datasets.

Ground-truth comparison. To verify the number of windows required to cover all pedestrians over an image, estimated by the theorem in Section III, this experiment determines how many windows are covering the pedestrian's bounding boxes according to their position given by the ground-truth as a function of the percentage of total detection windows sampled at random. Figure 2 shows that a random selection of 1.4% of detection windows is enough to detect more than 90% of pedestrians for both INRIA and ETHZ datasets. Note that these results show the maximum achievable recall if the detector provides perfect results, the next experiment evaluates the actual recall achieved by the PLS detector [17][1]

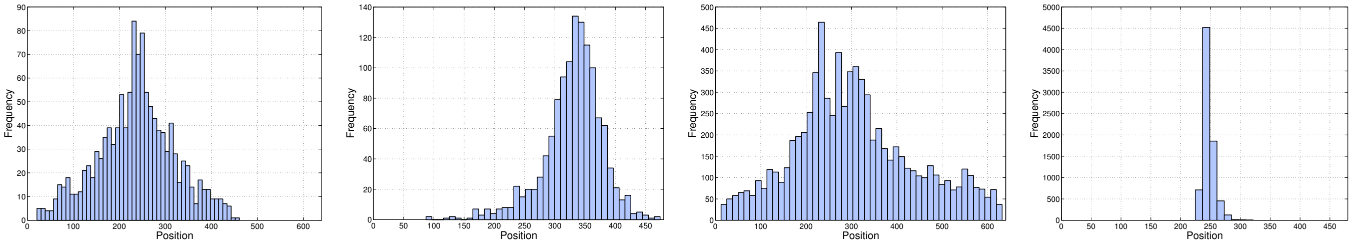
Pedestrian detector. After performing random filtering, it is possible to see that the selected windows practically do not miss any person in these datasets. However, these windows will be processed by a pedestrian detector, which may not obtain high accuracy due to these dislocated windows. To evaluate that, we feed the PLS detector with the random filtering output in order to classify these windows as containing pedestrians or not.

The results in Figure 4 show the recall obtained at one false positive per image (FPPI). Even after performing random filtering, the accuracy is still comparable to the original detection method, which considers 100% of the detection windows (no windows are discarded). However, to achieve the same results, the number of selected detection windows had to be larger than the result achieved by the ground truth experiment. This indicates that, even though the correct detection windows have been selected, the detector is not providing high responses for all the correct windows.

Distribution of objects. This work considers that pedestrians are uniformly distributed over an image. However, this is not necessarily true. This assumption may lead to more erroneous sampling or inefficient sampling.

We build histograms of the x- and y-coordinates of the

[1] For this work, the PLS detector was executed with a single stage. Therefore, it is not the same version evaluated in [17].



(a) INRIA (x-axis)

(b) INRIA (y-axis)

(c) ETHZ (x-axis)

(d) ETHZ (y-axis)

Fig. 3: Histogram of the distribution of the pedestrian according to the image coordinates in the x and y axes.

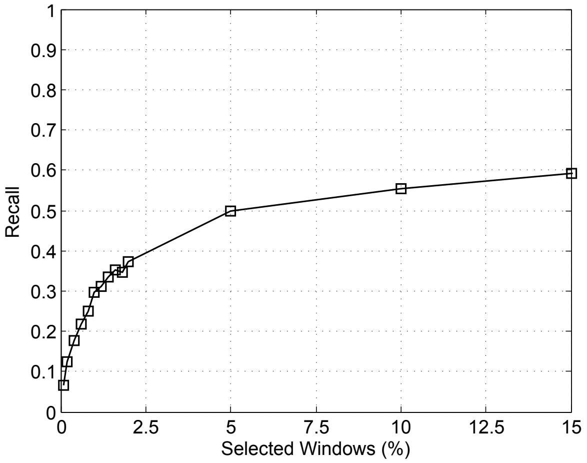


Fig. 4: Recall at 1 FPPI achieved when the selected detection windows are fed to the PLS detector. The execution of the detector for 100% of the detection windows (without selection) achieved a recall of 0.612.

pedestrians' centroids for every image. These histograms were collected from INRIA and ETHZ (sequence #1). According to the histograms displayed in Figure 3, it is clear that the distribution of the pedestrians in the frames is not uniform. We intend to study such characteristic in future works aiming at reducing even more the number of detection windows that have to be randomly sampled to detect all pedestrians in the image.

V. Conclusions and Future Works

This work proposed a detection optimization based on random filtering to discard a large number of detection windows and therefore reduce the computational cost. Our experimental evaluation showed that accurate results may be achieved even when a large number of detection windows are discarded.

As future works, we intend to finish evaluating this methodology in the ETHZ dataset. In addition, we will exploit the pedestrians' distribution in a dataset and measure the gain of this approach and, supported by the results of the experiment considering the ground truth and the real detector, we intend to devise a technique to improve the location of the sampled detection windows.

Acknowledgments

The authors are grateful to FAPEMIG, CAPES and CNPq funding agencies.

References

[1] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of Pedestrian Detection for Advanced Driver Assistance Systems," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 2010.

[2] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[3] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[4] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 1, pp. 36–51, 2008.

[5] Viola and Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2001.

[6] Q. Zhu, S. Avidan, M. chen Yeh, and K. ting Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2006, pp. 1491–1498.

[7] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2008.

[8] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *IEEE Intl. Conference on Computer Vision*, 2011, pp. 1028–1035.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[10] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances In Neural Information Processing Systems*, 2007, pp. 545–552.

[11] J. G. d. Silva Filho, L. Schnitman, and L. R. d. Oliveira, "Multi-scale spectral residual analysis to speed up image object detection," in *Brazilian Symposium on Computer Graphics and Image Processing*, 2012.

[12] C. Wojek, G. Dorkó, A. Schulz, and B. Schiele, "Sliding-Windows for Rapid Object Class Localization: A Parallel Technique," in *Proceedings of the 30th DAGM symposium on Pattern Recognition*. Springer-Verlag, 2008, pp. 71–81.

[13] L. Zhang and R. Nevatia, "Efficient scan-window based object detection using GPGPU," in *IEEE Computer Vision and Pattern Recognition Workshops*, 2008.

[14] I. Masaki, B. K. Horn, B. Bilgiç *et al.*, "Fast human detection with cascaded ensembles," Ph.D. dissertation, Massachusetts Institute of Technology, 2010.

[15] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization and beyond*, 2002.

[16] A. Ess, B. Leibe, and L. V. Gool, "Depth and appearance for mobile scene analysis," in *IEEE Intl. Conference on Computer Vision*, 2007.

[17] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis, "Human Detection Using Partial Least Squares Analysis," in *IEEE Intl. Conference on Computer Vision*, 2009.