A Late Fusion Approach to Combine Multiple Pedestrian Detectors

Artur Jordão, Jessica Sena de Souza, William Robson Schwartz
Smart Surveillance Interest Group, Computer Science Department
Universidade Federal de Minas Gerais, Minas Gerais, Brazil

***Abstract*—Pedestrian detection is a well-known problem in Computer Vision. To improve detection, several feature descriptors have been proposed and combined. However, there are cases where the most powerful features fail to discriminate between false positives similar to the human body structure and actual true positives, which is a critical problem for applications such as surveillance, driving assistance and robotics. To address this issue, we propose a novel approach to combine results of distinct pedestrian detectors by reinforcing the human hypothesis. The method is able to reduce the confidence of the false positives due to the lack of spatial consensus when multiple detectors are considered. Our experimental validation, performed on three pedestrian detection benchmarks, INRIA person, ETH and Caltech pedestrian dataset, demonstrates that the proposed approach, referred to as Spatial Consensus (SC), outperforms the state-of-the-art on INRIA and ETH datasets and achieves comparable results on the Caltech dataset.**

I. Introduction

Pedestrian detection has been an active research topic in Computer Vision, mostly because of its direct applications in surveillance, transit safety and robotics [1]. However, this task presents many challenges, such as occlusion, distinct illumination conditions and variance in human appearance.

Some works have shown that the majority of recent efforts in pedestrian detection can be attributed to the development of feature descriptors and evidences suggest that this trend will continue [1], [2]. In addition, several works show that feature combination creates a more powerful descriptor, improving detection [3], [4].

Dollár et al. [3] proposed the Integral Channel Features (ICF), where several features are extracted using integral images. Even though simple, ICF achieves accurate results for object detection and it has been the most employed feature in recent pedestrian detectors [1]. Aiming at finding human with high discriminative power, Javier et al. [4] extracted the HOG+LBP features of random patches inside a template window and achieved a detector robust to partial occlusions.

Besides the development of feature descriptors, previous works have also used high level information to refine the detections [1], [5]. Benenson et al. [1] demonstrated that extra information, such as context and optical flow, may improve pedestrian detection performance. Jiang and Ma [5] fused two detectors employing a novel non-maximum suppression technique called *weighted-NMS*, based on the hypothesis that an object is more likely to be a real object when it is detected by two different models (e.g., different pedestrian detection approaches). On the other hand, an object detected by a single model is more likely to be a false positive, as illustrated on the second column of Figure 1.

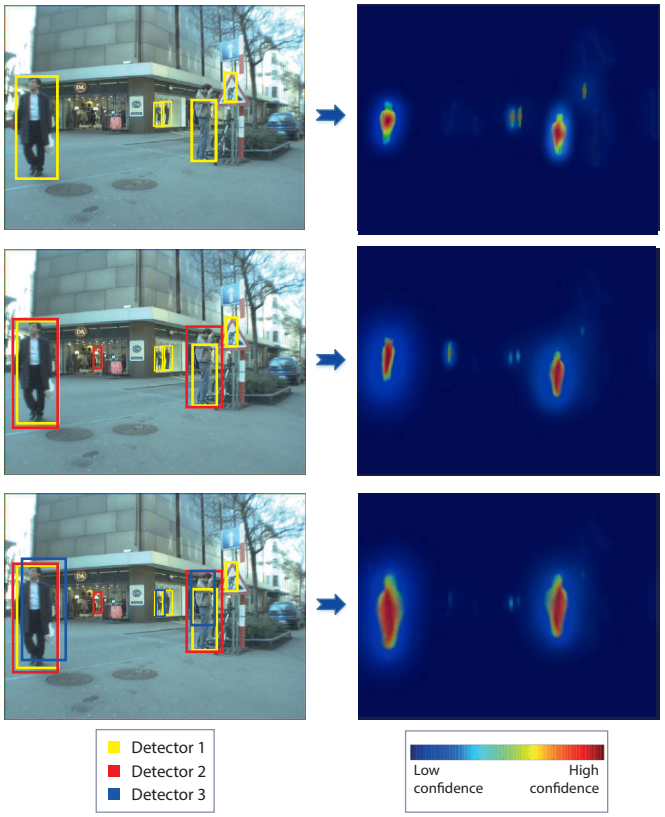


Fig. 1. Detection results and their respective heat map. On the top image, only one detector is being used to generate the heat map, but in the second and third images, two and three detectors, respectively, are used to generate the heat map. Each bounding box color represents the results of a distinct detector. As can be noticed, the addition of more detectors reduces the confidence of false positives with similar human body structure and reinforces the pedestrian hypothesis (best visualized in color).

The idea proposed by Jiang and Ma [5] is the following. If two detectors find the same object, given a determined overlapping area, the window with lower response is discarded and its confidence multiplied by a weight is added to the kept window. This is powerful because in the event of a true positive, the discarded window helps to increase the confidence of the kept one, while in the case of a false positive, it contributes to decrease the confidence. However, when the windows do not overlap, their method keeps both, which might

Detector	Feat. Type	Classifier	Occlusion Handled	training
SpatialPolling [6]	Multiple	pAUCBoost	-	INRIA/Caltech
S.Tokens [7]	ICF	Adaboost	-	INRIA+
Roerei [8]	ICF	AdaBoost	-	INRIA
Franken [9]	ICF	AdaBoost	✓	INRIA
LDCF [10]	ICF	AdaBoost	-	Caltech
I.Haar [11]	ICF	AdaBoost	-	INRIA/Caltech
SCCPriors [12]	ICF	AdaBoost	✓	INRIA/Caltech
NAMC [13]	ICF	AdaBoost	-	INRIA/Caltech
R.Forest [4]	HOG+LBP	D.Forest	✓	INRIA/Caltech
W.Channels [14]	WordChannels	AdaBoost	-	INRIA/Caltech
V.Fast [15]	ICF	AdaBoost	-	INRIA

TABLE I
OVERVIEW OF STATE-OF-THE-ART DETECTORS ON INRIA DATASET, SORTED BY LOG-AVERAGE MISS-RATE. TRAINING COLUMN: INRIA/CALTECH MODEL TRAINED USING INRIA AND CALTECH DATASETS; INRIA+ MODEL TRAINED USING INRIA DATASET WITH ADDITIONAL DATA.

increase the number of false positives.

Aiming at tackling the aforementioned limitation, we propose a novel late fusion method called *Spatial Consensus (SC)* to combine multiple detectors into a single detector named *root detector*. The root detector is composed of *virtual windows* generated by aggregating detection windows provided by multiple detectors, according to their locations and responses. The development of our method presents two main contributions: (1) we show that responses coming from extra detectors provide a strong cue to improve the detection since the spatial consensus for true positives increases their responses while the lack of spatial consensus for false positives helps to decrease its confidence (see the third row of Figure 1), and (2) we demonstrate that when dealing with the late fusion of multiple detectors, it is more appropriate to employ the proposed approach than the suppression procedure weighted-NMS described in [5].

To compare our approach with the *weighted-NMS* algorithm [5], we adapt the latter to enable the use of multiple detectors. According to the experimental results, our method reduces the false positive rate while keeping a high accuracy, which is desirable in applications such as surveillance and robotics [16]. In addition, the proposed method outperforms the state-of-the-art on the INRIA and ETH datasets with a log-average miss rate of 7.95% and 33.64%, respectively, and achieves comparable results on the Caltech dataset, with a log-average miss rate of 19.86%.

II. RELATED WORK

Following the work of Benenson et al. [1], we synthesize the main features of each detector instead of discussing each one individually. According to Table I, the majority of detectors employ Integral Channel Features (ICF) [3] combined with Adaboost algorithm [17], generally using decision stumps as weak classifier [15], [8], [13], [7], [9], [10], [12], [11]. The ICF-based detectors have the advantage of being extremely fast. The main differences among them are the way the features are preprocessed before classification, the training set used and the ability to handle partial occlusions.

Besides the ICF-based detectors, other approaches have achieved state-of-the-art results. Marin et al. [4] employed the traditional HOG+LBP [18] to describe human regions with high discriminative power by using random forests as classifier. Furthermore, combining HOG+LBP and LUV color channels, Costea et al. [14] generated high level visual words named *word channels*, allowing detection of pedestrians with different sizes considering a single scale image, which considerably reduces the computational cost.

Another line of research to improve pedestrian detection is the use of high level information regarding the objects in the scene. Such information can be obtained with the raw response of a single detector [19] or by combining distinct detectors [20]. Schwartz et al. [19] proposed an approach to learn a classifier using the raw responses of a general pedestrian detector. On the other hand, Li et.al [20] combined several pre-trained general object detectors, aiming at producing a more powerful image representation.

The successful results of approaches such as in [19], [20] rely on the hypothesis that regions containing a pedestrian have a strong concentration of high responses, different from false positive regions, where the responses present a large variance (low and high responses). Similar to [19], [20], our work captures additional information provided by a set of detectors through a late fusion approach. However, it is simpler and presents low computational cost since it does not require the employment of machine learning techniques.

Combination of results obtained by multiple detectors has also been explored for pedestrian detection [5], [21]. Ouyang and Wang [21] proposed a method to combine multiple detectors into a single detector to address the problem of groups of people focusing on cases where traditional sliding window approaches tend to fail. Similarly, Jiang and Ma [5] combined multiple detectors via a proposed *weighted-NMS* algorithm which, in contrast to the traditional non-maximum suppression algorithms, does not simply discard the window with the lowest score in the intersection, but uses the score to weight the kept window. On the other hand, the method proposed in this work targets on weight the windows of a single detector, ensuring that false positive windows provided by other detectors are not inserted.

III. Proposed Method

This section first describes the procedure to generate the root pedestrian detector needed to execute the SC algorithm. Then, we describe the steps of our proposed algorithm to combine responses of multiple detectors.

The idea behind building a root detector is to increase the flexibility of the algorithm – this way, we do not need specify a particular pedestrian detector as the input to the SC algorithm (see Algorithm 1). To generate windows for the root detector (det_{root}), let us consider the set of detectors $\{det_j\}_{j=1}^{n}$. For a detection window $w_i^j \in det_j$ with dimensions $(x, y, width, height)$, we search for overlapping windows in the remaining detectors $(w_i^l, l = 1, 2, ..., k)$ to create a set of windows that will be used to generate a single window belonging to the det_{root} using

$$w_i^{root} = \frac{1}{k} \sum_{l=1}^{k} w_i^l, \tag{1}$$

where k is the number of overlapping windows to the window w_i^j. Then, we assign a constant C (for instance, $C = 1$) to this novel window. This constant contains the score of this window and its value will be updated after executing the SC algorithm.

Now, we explain the use of responses coming from these detectors to weight the scores of a detection, giving more confidence to candidate windows that really belong to a pedestrian (our hypothesis is that regions containing pedestrians have a dense concentration of detection windows from multiple detectors converging to a spatial consensus, see Figure 1), whereas decreasing the confidence of the false positives.

The first issue to be solved when performing detector response combination (late fusion) is to normalize the output scores to the same range because different classifiers usually produce responses in their particular range. For instance, if the classifier used by the ith detector attributes a score in the range of $[-\infty, +\infty]$ to a given candidate window and the classifier of the jth detector attributes a score within $[0, 1]$, the scores cannot be combined directly. In this work, we employ the same procedure used in [5] to normalize the scores. The procedure steps are described as follows. First, we fix a set of recall points, e.g, $\{1, 0.9, 0.8, 0.7, ..., 0\}$. Then, for each detector, we collect the set of scores, τ, that achieve these recall points. Finally, we estimate a function that projects τ_j onto τ_i.

After normalizing the scores to the same range, we combine the candidate windows of different detectors as follows. Let det_{root} be the root detector from which the window scores will be weighted based on the detection windows of the remaining detectors in $\{det_j\}_{j=1}^{n}$. For each window $w_i \in det_{root}$, we search for windows $w_j \in det_j$ satisfying a specific overlap according to the *Jaccard coefficient* given by

$$J = \frac{\text{area}(w_i \cap w_j)}{\text{area}(w_i \cup w_j)}. \tag{2}$$

Then, we weight w_i in terms of

$$score(w_i) = score(w_i) + score(w_j) \times J. \tag{3}$$

Algorithm 1: Spatial Consensus (SC).

input : Candidate windows in det_{root}
output: Updated windows of det_{root}

```
for j ← 1 to n do
    project det_j score to a new score space;
    foreach w_i in det_root do
        foreach w_j in det_j do
            compute overlap using Equation 2;
            if overlap >= σ then
                update w_i score using Equation 3;
            end
        end
    end
end
```

The process described above is repeated n times, where n is the number of detectors besides the root detector. Algorithm 1 represents the aforementioned process.

Regarding the computational cost, the asymptotic complexity of our method is denoted by

$$O(p_{root} \times \sum_{j=1}^{n} p_j) = O(p_{root} \times p) = O(p^2),$$

where p_{root} is the number of candidate windows of det_{root}, p_j denotes the number of detection windows of the jth detector and p is the amount of all candidate windows in $\{det_j\}_{j=1}^{n}$. Similarly, the approach proposed by Jiang and Ma (weighted-NMS method) [5] presents complexity of $O(p \log p + p^2)$. Although both methods present a quadratic complexity, p is extremely small because the traditional non-maximum suppression is employed for each detector before presenting the candidate windows to Algorithm 1, which renders the computational cost of both our Spatial Consensus method and the baseline approach in [5] to be negligible when compared with the execution time of the individual pedestrian detectors.

Our approach differs from the weighted-NMS method [5] in the following aspect. Instead of inserting the candidate windows of all the detectors to be combined into a single set and performing weighted-NMS (see Section IV), we generate a root detector containing virtual windows and weight them by the responses of the overlapping windows from the actual detectors. In this way, we reduce the chance of adding errors by choosing a window that poorly covers the pedestrian, according to the ground-truth.

To illustrate the difference between weighted-NMS and SC algorithm, let us consider the scenario demonstrated in Figure 2, where the red and yellow windows present a jaccard coefficient of 0.8 and 0.7, respectively, regarding the black window. The suppression employed by the weighted-NMS algorithm would choose the red window (losing the pedestrian and generating a false positive and a false negative), since it presents higher score regarding the yellow window. Our algorithm would generate the black window (average window

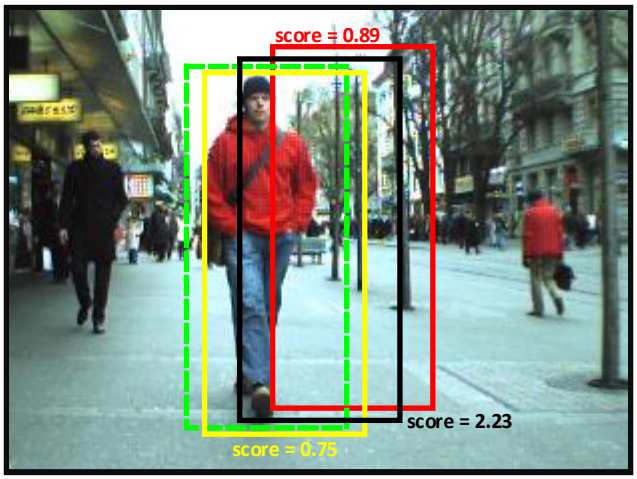


Fig. 2. Yellow and red boxes indicate the detection coming from det_j and the green box shows the ground-truth annotation. The weighted-NMS [5] will maintain the red box (false positive) because it is the window with higher score, leading to higher miss rate and reduced recall. Our method will create the black window, with a high confidence on the pedestrian (best visualized in color).

between the red and yellow windows) with score equal to one. Considering the red and yellow detectors as the first and second elements in $\{det_j\}_{j=1}^{n}$, in the first iteration of the SC algorithm the black window score will be updated to 1.712 $(1 + 0.89 * 0.8)$. Then, in the second iteration, its score will be updated to 2.23 $(1.712 + 0.75 * 0.7)$. As can be noticed, our method increased the confidence on the pedestrian besides providing a higher intersection with the ground-truth window, in green, whereas weighted-NMS lost the pedestrian.

IV. Experimental Results

To validate our results, we adopt the test protocols established in the Caltech pedestrian benchmark [22], where the results are reported on the log-average miss rate (lower values are better). A detailed discussion regarding this protocol of evaluation can be found in [22], [2]. We use the code available in the toolbox[1] of this benchmark to perform the evaluation.

Preparing the input detectors. Initially, we need to define a set of detectors $\{det_j\}_{j=1}^{n}$ to compose and weight our root detector (as explained in Section III). Due to the large number of pedestrian detectors currently available, there are many options to determine $\{det_j\}_{j=1}^{n}$ [1], [2]. In this work, we define these detectors as the top eleven best ranked pedestrian detectors on the INRIA person dataset (Table I). The best ranked detector, the SpatialPooling [6], was utilized to perform the step of score normalization.

The columns of Table II show the detectors used in $\{det_j\}_{j=1}^{n}$ in the INRIA dataset. In our proposed algorithm, each detector in det_j is utilized sequentially to produce and weight the det_{root}. Note that the order of the set members does not affect the result since all detectors of $\{det_j\}_{j=1}^{n}$ must be evaluated to compose and weight a window of det_{root}.

[1] www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

At the score calibration step, we use the INRIA person dataset to acquire the set of scores τ. Next, to map the $\{det_j\}_{j=1}^{n}$ score to the Spatial Pooling score, we consider a linear regression. From the scatter plot between $\tau_{SpatialPooling} \times \tau_j$, we observed that a linear regression is a suitable choice to perform this mapping.

We preferred not to implement some detectors since we would hardly obtain the correct results due to parameter setup. Therefore, throughout of the experiments we will use the results provided by authors and this fact forced us to use INRIA test to calibrate the scores only to one of our experiments (only to produce Table II). However, once the score are calibrated, we use the estimated regression on the other datasets.

It is important to mention that before combining the detectors by Algorithm 1 or by the weighted-NMS algorithm [5], we assume that all detectors performed non-maximum suppression (NMS) individually. This initial NMS is performed to suppress overlapping detections from the same detector and that is essential to reduce the number of candidate windows since it will influence the running time of both algorithms.

Weighted-NMS baseline. The method proposed by Jiang and Ma [5] can be described in four main steps. First of all, the detection window responses of the combined detectors are normalized to the same score range. Then, the windows of both detectors are inserted in a set U. Afterwards, U is sorted in descending order according to their scores. Finally, when a window at position i of U presents overlap higher than the threshold (σ) with a window at position j of U ($j > i + 1$), the NMS process is applied and the window with the lower score is discarded and its score contributes to the kept window (similar to step 7 of Algorithm 1).

Since the method proposed in [5] accepts only two detectors as input, we insert all windows of the detectors to combine into U, which makes it handle multiple detectors simultaneously.

Spatial Consensus vs. weighted-NMS. In this experiment, aiming a fair comparison, we report the weighted-NMS results achieved by better combination of detectors in each dataset. To our approach, we report the results always using the ten detectors (as shown in Table II) available. In addition, using the INRIA Person dataset, we estimated the best thresholding σ applied on the *Jaccard coefficient*, where this value was 0.6 for both the methods, see Table II.

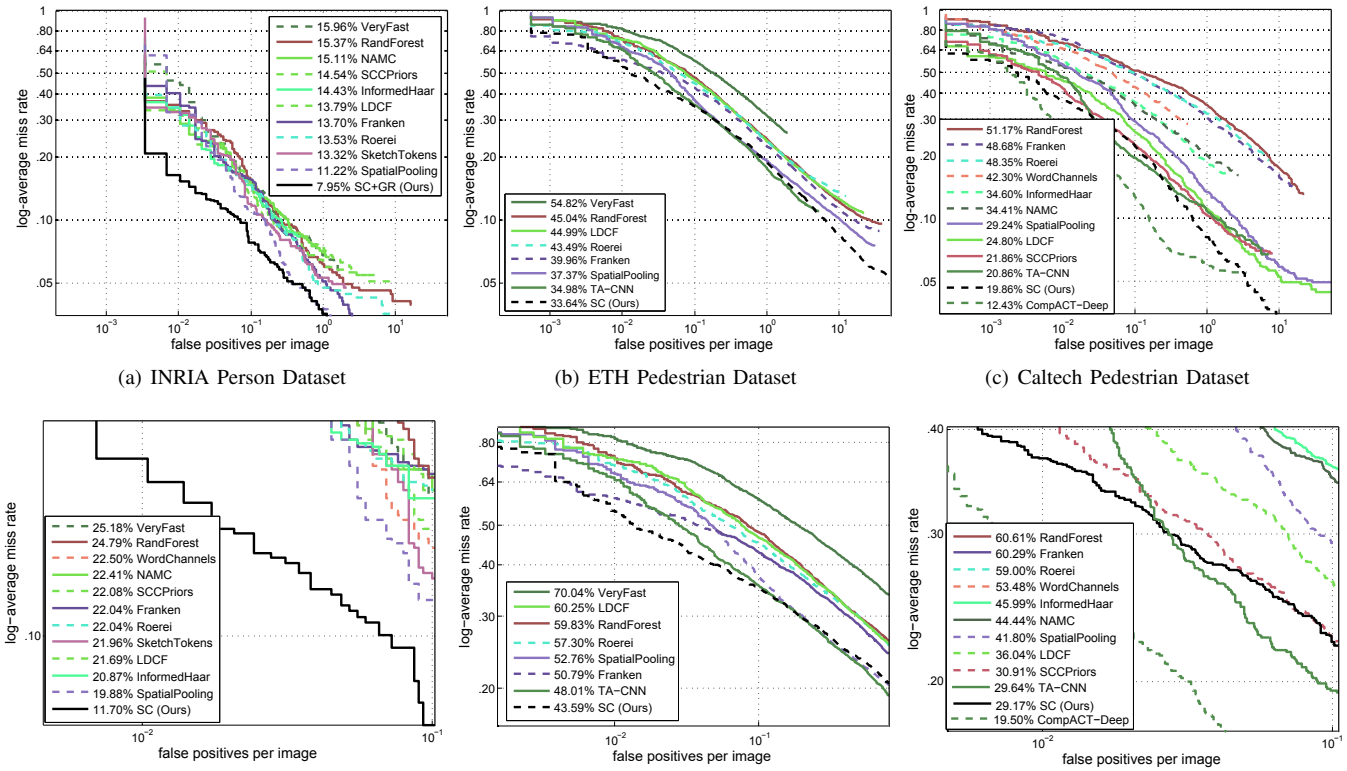
The weighted-NMS achieved the best results on the INRIA dataset when two detectors are added, Sketch Tokens [7] and Roerei [8], outperforming the state-of-the-art by 1.48 percentage points (p.p.), while the best result of our approach is achieved adding all the detectors, outperforming the state-of-the-art in 3.27 p.p. (log-average miss rate of 7.95%).

On the ETH dataset, the weighted-NMS method achieved its best result, 35.19%, by combining Roerei and Franken [9] detectors. This combination was not enough to outperform the TA-CNN [23] (state-of-the-art on this dataset with 34.98%). On the other hand, our approach outperforms the state-of-the-art result in 1.34 p.p. (log-average miss rate of 33.64%).

σ		S.Tokens	Roerei	Franken	LDCF	I.Haar	SCCPriors	NAMC	R.Forest	W. Channels	V.Fast
0.5	SC (Ours)		10.09%	10.11%	10.03%	9.53%	9.70%	10.13%	9.47%	8.63%	**8.25**%
	W-NMS	**10.60**%	11.22%	12.91%	12.42%	12.48%	12.37%	12.38%	14.70%	14.81%	14.11%
0.6	SC (Ours)		9.98%	9.45%	9.11%	9.00%	9.35%	10.07%	9.13%	8.46%	**7.95%**
	W-NMS	10.12%	**9.74**%	11.75%	11.84%	12.72%	13.48%	13.65%	16.11%	14.81%	14.11%
0.7	SC (Ours)		10.69%	9.41%	9.63%	9.42%	9.55%	9.77%	9.42%	8.71%	**8.21**%
	W-NMS	11.37%	**10.01**%	13.58%	14.81%	15.59%	16.14%	17.79%	24.88%	14.81%	14.11%

TABLE II

INRIA PERSON DETECTORS ACCUMULATION. THE INITIALS SC REFERS TO OUR PROPOSED METHOD AND THE INITIALS W-NMS REFERS TO OUR BASELINE THE WEIGHTED-NMS [5]. THE RESULTS ARE MEASURED IN LOG-AVERAGE MISS-RATE (LOWER IS BETTER). FROM THE LEFT TO THE RIGHT, EACH COLUMN k DENOTES THE ADDITION OF THE RESPECTIVE DETECTOR INTO $\{det_j\}_{j=1}^{n}$.



(a) INRIA Person Dataset (b) ETH Pedestrian Dataset (c) Caltech Pedestrian Dataset

(d) INRIA Person Dataset (e) ETH Pedestrian Dataset (f) Caltech Pedestrian Dataset

Fig. 3. Comparison of our proposed approach with the state-of-the-art. The first row reports the results using the log-average miss-rate of 10^{-2} to 10^{0} (standard protocol). The second row reports the results using the area of 10^{-2} to 10^{-1}

The best result of the weighted-NMS on the Caltech dataset was achieved combining the Roerei and Franken detectors. With this combination the log-average miss rate was of 40.54%. On the contrary, our SC method outperformed the baseline in 20.68 p.p. (log-average miss rate of 19.86%).

According to the results, we conclude that the proposed method is more suitable to perform fusion between multiple detectors than the weighted-NMS [5].

Influence of a less accurate detector. To evaluate the robustness of our method to the addition of a detector with high false positive rate, we introduced the HOG detector [24] (log-average miss rate of 46%) right after the V. Fast [15] on the INRIA person dataset. When it was inserted into $\{det_j\}_{j=1}^{n}$, the log-average miss rate achieved by our method was from 7.95% to 10.95% and the weighted-NMS the result was from 14.11% to 16.78%, demonstrating that our algorithm is more robust to less accurate detectors.

Comparison with the state-of-the-art. In this experiment, we compare the results of the proposed *Spatial Consensus* algorithm with state-of-the-art methods. To perform a fair comparison, we considered the results reported by the authors in their works.

Figures 3(a) and (b) show that our algorithm outperforms the state-of-the-art on the INRIA and ETH datasets achieving log-average miss-rate of 7.95% and 33.64%, respectively. Furthermore, Figure 3(c) shows that our method achieves significant results on the Caltech dataset with log-average miss rate of 19.86%.

An important goal of pedestrian detection is to significantly minimize false alarms for applications such as video surveillance in which they may cause damage to environment as well to humans [16]. To indicate that our method is suitable for the requirement of very low false positive rates, we report our results using the area under curve from 10^{-2} to 10^{-1} (values where the false positive rates are extremely small). Figure 3 (d), (e), and (f) show these results. As can be noticed, our method further enhances the detection accuracy, demonstrating to be appropriate to applications that need to operate at very low false positive rates.

Time issues. As described in Section III, the complexity of our method is equal to the weighted-NMS. Although quadratic, both methods run in real time since the traditional NMS is performed for each individual detector before starting the algorithms (see Section IV). Besides, the values of p_{root} and p are corresponding to the number of pedestrians at the scene, which is usually very low. To verify that these values are extremely small, we collected the average of people per image in the INRIA person and the ETH ($seq\#2$) datasets. The values are 3.3 and 43.6, respectively (not large enough to impact the computational time of our algorithm).

Since the values of p are small, our approach is able to run in real time. To show that, we computed the time average to execute of the SC on an image 640×480 pixels, using 10 detectors to compose $\{det_j\}_{j=1}^{n}$ and without any parallelization technique. The SC runs in 67 milliseconds, on average (this experiment was executed 10 times). Additionally, the most recent survey of computation cost at the detection pedestrian [16], showed that the faster detector presenting high accuracy is able to process 15 frames per second on a GPU [16]. Therefore, we conclude that our method is able to improve the detection results and could be fast to execute, even though our algorithm requires results of individual detectors.

V. Conclusions and future works

This work presented an efficient and low cost method, called *Spatial Consensus (SC)*, to combine multiple pedestrian detectors. The method focuses on the idea of using the responses coming from multiple detectors to reinforce more consistent human hypothesis whereas reducing the confidence of the false positives. The proposed method outperformed the state-of-the-art in two pedestrian detection benchmarks and achieved comparable results on the challenging Caltech dataset. As future work, we intend to apply the proposed algorithm to others object categories based on sliding window and use the responses coming from multiple detectors as raw features, similar to [19].

Acknowledgments

The authors would like to thank the Brazilian National Research Council – CNPq (Grant #477457/2013-4), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00025-15) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project).

References

[1] R. Benenson, M. Omran, J. Hosang, , and B. Schiele, "Ten years of pedestrian detection, what have we learned?," in *ECCV*, 2014.
[2] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: An evaluation of the state of the art," *In PAMI*, 2012.
[3] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie, "Integral channel features.," in *BMVC*, 2009.
[4] Javier Marín, David Vázquez, Antonio M. López, Jaume Amores, and Bastian Leibe, "Random forests of local experts for pedestrian detection," in *ICCV*, 2013.
[5] Yunsheng Jiang and Jinwen Ma, "Combination features and models for human detection," in *CVPR*, 2015.
[6] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel, "Strengthening the effectiveness of pedestrian detection with spatially pooled features," in *ECCV*, 2014.
[7] Joseph J. Lim, C. Lawrence Zitnick, and Piotr Dollár, "Sketch tokens: A learned mid-level representation for contour and object detection," in *CVPR*, 2013.
[8] Rodrigo Benenson, Markus Mathias, Tinne Tuytelaars, and Luc J. Van Gool, "Seeking the strongest rigid detector," in *CVPR*, 2013.
[9] Markus Mathias, Rodrigo Benenson, Radu Timofte, and Luc J. Van Gool, "Handling occlusions with franken-classifiers," in *ICCV*, 2013.
[10] Woonhyun Nam, Piotr Dollár, and Joon Hee Han, "Local decorrelation for improved pedestrian detection," in *NIPS*, 2014.
[11] Shanshan Zhang, Christian Bauckhage, and Armin B. Cremers, "Informed haar-like features improve pedestrian detection," in *CVPR*, 2014.
[12] Yi Yang, Zhenhua Wang, and Fuchao Wu, "Exploring prior knowledge for pedestrian detection," in *BMVC*, 2015.
[13] M. Ciuc C. Toca and C. Patrascu, "Normalized autobinomial markov channels for pedestrian detection," in *BMVC*, 2015.
[14] Arthur Daniel Costea and Sergiu Nedevschi, "Word channel based multiscale pedestrian detection without image resizing and using only one classifier," in *CVPR*, 2014.
[15] Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc J. Van Gool, "Pedestrian detection at 100 frames per second," in *CVPR*, 2012.
[16] Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson, "Real-time pedestrian detection with deep network cascades," in *BMVC*, 2015.
[17] Robert E. Schapire, "A brief introduction to boosting," in *IJCAI*, 1999.
[18] Xiaoyu Wang, Tony X. Han, and Shuicheng Yan, "An HOG-LBP human detector with partial occlusion handling," in *ICCV*, 2009.
[19] William Robson Schwartz, Larry S. Davis, and Hélio Pedrini, "Local response context applied to pedestrian detection," in *CIARP 2011*.
[20] Li-Jia Li, Hao Su, Eric P. Xing, and Fei-Fei Li, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *NIPS*, 2010.
[21] Wanli Ouyang and Xiaogang Wang, "Single-pedestrian detection aided by multi-pedestrian detection," in *CVPR*, 2013.
[22] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: A benchmark," in *CVPR*, 2009.
[23] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Pedestrian detection aided by deep learning semantic tasks," in *CVPR*, 2015.
[24] Navneet Dalal and Bill Triggs, "Histograms of Oriented Gradients for Human Detection," in *CVPR*, 2005.