Supplementary material for: Statistical and Spatial Consensus Collection for Detector Adaptation

Enver Sangineto

DISI, University of Trento, Italy*

1 Introduction

The aim of this supplementary material is to provide further experimental details which were not included in the main article due to space limits (Sec. 2) and to give more details concerning the choice of our proposed RANSAC-like optimization strategy (Sec. 3).

Specifically, in Sec. 2 we show further results, obtained with the CUHK Square Test dataset only in which we partially corrected the annotations provided by the authors of [10], adding some pedestrian bounding boxes missing in the original ground truth.

As mentioned in the main article (Sec. 6), we used for our experiments the MIT Traffic and the CUHK Square datasets adopted by Wang et al. in [10]. We used the videos downloaded from [8] and [2], respectively. We followed the instructions contained in the README files available at [8] and [2], in which it is suggested to resize into 6 times the frames of the MIT Traffic dataset and 2 times the frames of the CUHK Square dataset. The final frame resolutions are: 2880 by 4320 (MIT Traffic) and 1152 by 1440 (CUHK Square). The results are shown in Fig. 4 in the main article and *they have been obtained using the original ground truth downloaded together with the videos and used by the other authors we compare with*. By contrast, in Sec. 2 of this Supplementary Material we show further results on one of the videos in which we have partially corrected some errors in the original ground truth data.

Finally we point out that the comparison with other authors and other application settings is not always easy. For instance, the very low resolution videos we adopted make the pedestrian detection task very challenging for "standard" generic pedestrian detectors. As mentioned in Sec. 3 of the main article, more sophisticated detectors such as [3] completely failed when tested with our videos because of the low resolution and could not be used as a baseline for the extraction of our candidate positive samples (Wang et al. in [10] also tested the Felzenszwalb's detector [3] on the same videos obtaining completely unreliable results). Conversely, the authors of [9] used two benchmarks for their experiments, one of which is the Mind's Eye's Dataset Year 2 which is not publicly available at the moment. Anyway, the target videos used in [9] have a much

* Enver.Sangineto@unitn.it

higher image resolution than the ones used in our experiments, a resolution which is much closer to our source dataset (the INRIA dataset), thus we believe that these benchmarks are less challenging than the MIT Traffic and the CHUCK Square data and, hence, that our approach should work pretty well on that videos.

On the other hand, In Sec. 3 we better motivate our choice of using a stochastic RANSAC-like strategy to deal with the large number of false positives produced by the generic detector in the initial pedestrian candidate set T. In that section we examine other optimization approaches and we show that they cannot be applied to our problem.

2 Further Results on the CUHK Square Test Set

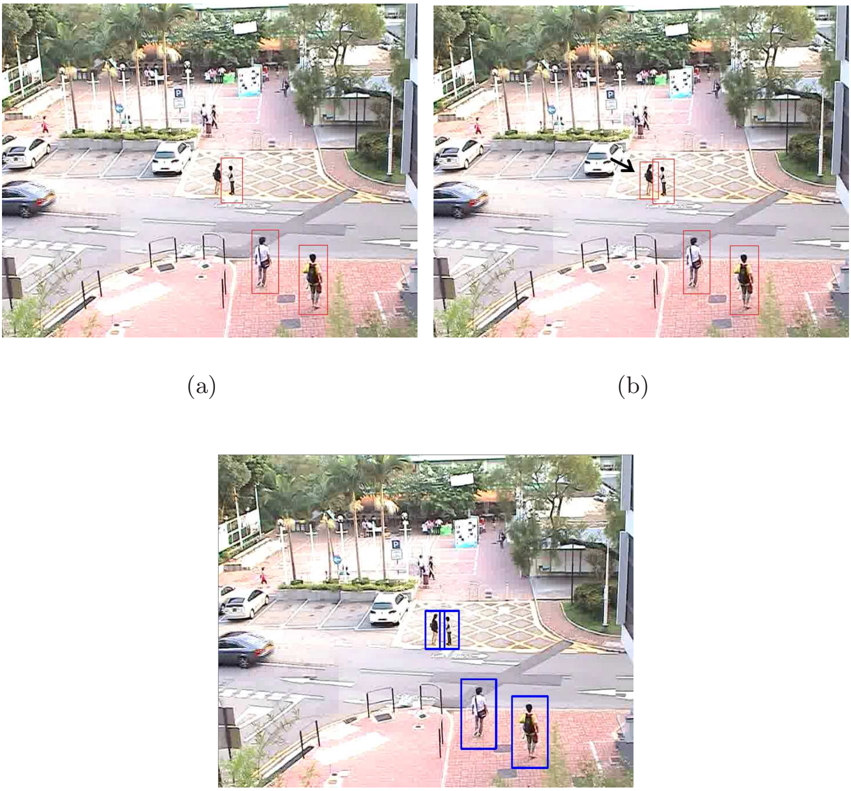
We noticed that some pedestrians in both the CUHK Square Train and the CUHK Square Test datasets are missing. Most of them are very small, but some others are larger. Some examples are shown in Figure 1(a).

For this reason we decided to augment the ground truth data with missing pedestrians. *Note that* ***all*** *the experiments presented in the main article (Figures 1-3 and Tables 1-2 in the article) have been obtained using only the ground truth data used in [10] and downloaded from [8] and [2].* However, for completeness, we show here the results obtained with a more complete ground truth.

The criteria we adopted for adding new pedestrian bounding boxes are the following. We used only the CUHK Square Test dataset and we corrected the annotations of all the 100 frames (when necessary). We added annotations for all those pedestrians whose resulting bounding box size ("padding" included) in the final 1152 by 1440 enlarged resolution frame was at least 64 pixels width. This choice was done because 64 × 128 is the minimum size sliding window for HOG based systems. However, occasionally some added pedestrians have a hight lower than 128 pixels. We also added truncated pedestrians but only if at least the full upper body is visible. An example of a new annotation is shown in Figure 1(b).

Using this new ground truth, we obtained the results shown in Figure 2. The red dash-dotted line with diamonds (SSCC-new GT) shows the results of our system with the new ground truth. All the other ROC curves in the same figure *have been obtained using the original ground truth* [8, 2], thus a direct comparison is *not* possible. However, we plotted the other ROC curves to allow the reader to get an intuitive idea of the difference depending from the annotations.

Specifically, in the high precision region of the graph (FPPI less than 1), (SSCC-new GT) shows better results than the ones we obtained using the original ground truth (SSCC). This is probably due to the fact that some true pedestrians, correctly detected by our system, are wrongly penalized as they were false positives using the original ground truth (see Figure 1(c)). Conversely, in the high recall region of the graph (FPPI greater than 1), our method obtained better results with the original ground truth. This is probably due to the fact that some of the pedestrian bounding boxes we added are very small or the con-



(a)

(b)

(c)

Fig. 1. Ground truth comparison. (a) The original ground truth [2] for an example frame of the CUHK Test dataset. (b) Our ground truth. The arrow shows the added annotation. Note that we did not added any annotation for the very small pedestrians in the background. (c) The results of our system with this image and the new ground truth. In this case there is an improvement because the leftmost pedestrian, whose bounding box is missed in the original ground truth, is wrongly counted as a false positive in the experiments performed in the article. However, all the other systems tested on the same datasets obtained their results using the same, incomplete ground truth, thus supposedly our system is not penalized more than the others. Unfortunately, since we do not have the code of the other approaches we could not repeat the experiments with our ground truth.

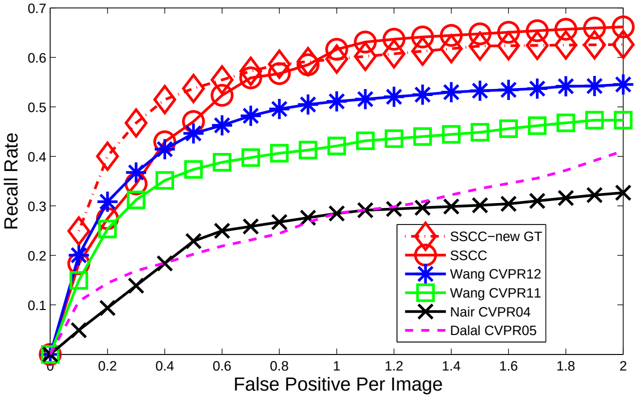


Fig. 2. Experiments on the CUHK Square Test dataset. The red dash-dotted line with diamonds (SSCC-new GT) shows the only results obtained using our ground truth. (SSCC-new GT) and (SSCC) are exactly the same system (i.e. the approach we proposed in the main paper), with exactly the same parameter setting and the same classifier ensemble (which were *not* retrained), being the used ground truth at testing time the only difference between these two curves.

tained human body is highly occluded or truncated and our system is not able to detect these difficult instances.

We can provide our new ground truth for the CUHK Square Test set to people possibly interested in testing their approaches on the same data.

3 Optimization Approaches for Subset Selection

The main problem addressed in our paper, i.e., how to select candidate pedestrian samples for training a target classifier directly on target data but without using new labels, is formulated in the Eq.s (3) and (4) of the main paper that we rewrite below for clarity.

$$T_G = \arg \min_{T_i \subseteq T} L(C_{T_i}, S), \tag{1}$$

subject to:

$$C_{T_i} = \arg \min_{C \in \mathcal{C}} \mathcal{R}(C) + \theta \lambda (T_i, N_i). \tag{2}$$

Since $L(\cdot) \in [0, 1]$ (see Eq. (6) of the main article), using $L'(\cdot) = 1 - L(\cdot)$, Eq. (1) above can be transformed into:

$$T_G = \arg \max_{T_i \subseteq T} L'(C_{T_i}, S). \tag{3}$$

Both $L(\cdot)$ and $L'(\cdot)$ are set functions (because they depend on the set T_i) and the set function maximization problem is known to be NP-hard [5]. When the function to be maximized ($L'(\cdot)$) satisfies some specific assumptions (e.g., submodularity or adaptive submodularity), the corresponding optimization can be solved using some simple (but in some cases computationally very expensive) greedy algorithms [5–7]. However, our problem formulation (Eq.s (1)-(2) above) does not satisfy these properties and different strategies must be adopted. The aim of the following subsections is to show why submodularity-based greedy algorithms cannot be applied in our case and to justify our choice in favor of a RANSAC-like algorithm as presented in the main paper.

3.1 Submodularity

In [5] the authors report a few simple greedy algorithms which can be used in order to solve problems formulated as below:

$$\max_{A \subseteq V} f(A), \text{subject to some constraints on} A, \tag{4}$$

where A and V are sets. These greedy algorithms are guaranteed to reach a good approximation to the optimal solution of the NP-hard optimization problem. However, they cannot be applied to our case because $f(\cdot)$ needs to belong to the *submodular* function class and $L'(\cdot)$ is not submodular. We provide below an informal proof starting from the definition of submodularity [5].

Definition (Discrete derivative). For a set function $f : 2^V \rightarrow \mathbb{R}$, $A \subseteq V$ and $e \in V$, let $\Delta_f(e|A) := f(A \cup \{e\}) - f(A)$ be the *discrete derivative* of $f(\cdot)$ at A with respect to e.

Given the previous definition, the submodularity property of a set function can then be defined as follows.

Definition (Submodularity). A function $f : 2^V \rightarrow \mathbb{R}$ is *submodular* if for every $A \subseteq B \subseteq V$ and $e \in V \setminus B$ it holds that:

$$\Delta_f(e|A) \geq \Delta_f(e|B). \tag{5}$$

We show now why $L'(\cdot)$ is *not* submodular and, hence, greedy submodular optimization algorithms cannot be applied for solving Eq.s (1)-(2) above.

The reason for which L and L' are not submodular is their non-smooth dependence on the results of the optimization in Eq. (2). For instance, suppose that the class of classifiers $\mathcal{C}$ from which C_{T_i} is chosen is the linear Support Vector Machine classifiers, which is what we adopted in our experiments. Moreover, suppose that T_1, T_2 are two different training sets (random subsets of candidate pedestrians in our case), $T_1 \subset T_2$ and that H_1, H_2 are the decision hyperplanes respectively obtained minimizing Eq. (2) with respect to T_1 and T_2. Finally, suppose that sample $\mathbf{x}$ is such that $\mathbf{x} \notin T_2$.

Let $T_1' = T_1 \cup \{\mathbf{x}\}$ and $T_2' = T_2 \cup \{\mathbf{x}\}$ and H_1', H_2' be the corresponding decision hyperplanes obtained minimizing Eq. (2) with respect to T_1' and T_2' respectively. We can easily choose $\mathbf{x}$ far from H_1 and close to H_2 such that:

- Adding $\mathbf{x}$ to T_1, the resulting decision hyperplane H_1' does not change ($H_1' = H_1$).
- Adding $\mathbf{x}$ to T_2, the resulting decision hyperplane H_2' does change ($H_2' \neq H_2$) (e.g., $\mathbf{x}$ is a support vector of H_2').

Since $H_1' = H_1$, then, by definition, $C_{T_1} = C_{T_1'}$, hence $L'(C_{T_1'}, S) = L'(C_{T_1}, S)$. However, since $H_2' \neq H_2$, we have that $L'(C_{T_2'}, S) \neq L'(C_{T_2}, S)$ and, for some $\mathbf{x}$, we can have $L'(C_{T_2'}, S) > L'(C_{T_2}, S)$ (which happens when the generalization ability of H_2' is improved with respect to the generalization ability of H_2), which violates the definition in (5) (with: $A = T_1, B = T_2$ and $e = \mathbf{x}$).

3.2 Adaptive Submodularity and Noise

In [6] Krause and Golovin extend the submodularity framework to *adaptive submodularity* and propose some greedy algorithms which can be used to solve adaptive submodular optimization problems. An interesting application of adaptive submodular optimization is *active learning*, where one must adaptively select data points to label in order to maximize the performance of a classifier trained on the selected data points.

The key idea is to substitute the function $f(\cdot)$ in Eq. (4) with $f_{ave}(\pi)$, where π is a *policy* (e.g., a sample selection policy), the states (e.g., the labels) of the sample set are drawn from a given probability distribution Φ and the score of π, $f_{ave}(\pi)$, is computed as the expectation over Φ. The definition of adaptive submodularity is then obtained extending the definition of discrete derivative (see Sec. 3.1) taking into account the *expected* increment of $f_{ave}(\pi)$ over Φ. Finally, Krause and Golovin also extend the greedy algorithms for submodular function maximization to deal with adaptive submodular function optimization problems. These greedy algorithms can be computationally very expensive but the authors provide lower bounds to the optimal solution of the original optimization problem.

It is not completely clear to us if our score function $L'(\cdot)$ can be shown to be adaptive submodular. Anyway, it can be easily shown that it is not *adaptive monotone* [6] because our pedestrian candidate set T contains outliers and the algorithms proposed in [6] can only be applied to free-noise samples (i.e., adaptive monotone functions). For more details we refer the reader to [6].

The noise case is dealt with in [7], where Guillory and Bilmes propose a greedy algorithm for optimizing the set selection problem in active learning problems with label noise. However, this algorithm can be applied only to submodular functions and we showed in Sec. 3.1 that $L'(\cdot)$ is not submodular.

3.3 Our RANSAC-like Optimization Strategy

We are not aware of any other greedy optimization strategy which can be applied to the problem formulated as in Eq.s (1)-(2) (respectively, Eq.s (3)-(4) of the main article). As we showed in the two previous subsections of this Supplementary Material, submodular or adaptive monotone submodular function-based greedy optimization approaches cannot be applied because of the specific nature of our problem. For this reason we adopted a stochastic strategy conceptually similar to the well known RANSAC algorithm [4]. RANSAC techniques have been successfully adopted in many Computer Vision problems in more then 30 years and are usually considered very robust [1].

As explained in the main article, Eq.s (1)-(2) are solved transforming the set selection problem in a RANSAC-like procedure, where we build models (classifiers) using random subsets of target data and we verify the accuracy of each classifier on the source set S, using the implicit assumption that target and source domains are related. This is the main novelty we propose and we provided in the paper the experimental results which confirm its validity.

References

1. Choi, S., Kim, T., Yu, W.: Performance evaluation of RANSAC family. In: BMVC. pp. 1–12 (2009)
2. CUHK Square: The CUHK Square dataset., available at: `http://www.ee.cuhk.edu.hk/~xgwang/CUHKSquare.html`
3. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. on PAMI 32(9), 1627–1645 (2010)
4. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24(6), 381–395 (1981)
5. Golovin, D., Krause, A.: Tractability: Practical Approaches to Hard Problems (to appear), chap. Submodular Function Maximization
6. Golovin, D., Krause, A.: Adaptive submodularity: Theory and applications in active learning and stochastic optimization. Journal of Artificial Intelligence Research (JAIR) 42, 427–486 (2011)
7. Guillory, A., Bilmes, J.: Simultaneous learning and covering with adversarial noise. In: ICML. pp. 369–376 (2011)
8. MIT Traffic: The MIT Traffic dataset., available at: `http://www.ee.cuhk.edu.hk/~xgwang/MITtraffic.html`
9. Sharma, P., Nevatia, R.: Efficient detector adaptation for object detection in a video. In: CVPR. pp. 3254–3261 (2013)
10. Wang, M., Li, W., Wang, X.: Transferring a generic pedestrian detector towards specific scenes. In: CVPR. pp. 3274–3281 (2012)