Weakly Supervised Automatic Annotation of Pedestrian Bounding Boxes

David Vázquez[1], Jiaolong Xu[1], Sebastian Ramos[1], Antonio M. López[1,2] and Daniel Ponsa[1,2]

[1]Computer Vision Center [2]Dept. of Computer Science
Autonomous University of Barcelona
08193 Bellaterra, Barcelona, Spain
{dvazquez, jiaolong, sramosp, antonio, daniel}@cvc.uab.es

Abstract

Among the components of a pedestrian detector, its trained pedestrian classifier is crucial for achieving the desired performance. The initial task of the training process consists in collecting samples of pedestrians and background, which involves tiresome manual annotation of pedestrian bounding boxes (BBs). Thus, recent works have assessed the use of automatically collected samples from photo-realistic virtual worlds. However, learning from virtual-world samples and testing in real-world images may suffer the dataset shift problem. Accordingly, in this paper we assess an strategy to collect samples from the real world and retrain with them, thus avoiding the dataset shift, but in such a way that no BBs of real-world pedestrians have to be provided. In particular, we train a pedestrian classifier based on virtual-world samples (no human annotation required). Then, using such a classifier we collect pedestrian samples from real-world images by detection. After, a human oracle rejects the false detections efficiently (weak annotation). Finally, a new classifier is trained with the accepted detections. We show that this classifier is competitive with respect to the counterpart trained with samples collected by manually annotating hundreds of pedestrian BBs.

1. Introduction

The task of an image-based pedestrian detector consists in locating the pedestrians that a given image contains, *e.g.* by framing each one with a bounding box (BB). Such an ability is the core of different emerging applications like in the fields of surveillance and driver assistance. Pedestrian detection is a difficult task due to the variability in the appearance of both pedestrians (size, pose and clothes) and their surrounding environment (illumination and background). Not surprisingly, recent surveys [5, 10, 3] reveal image-based pedestrian detection as a very active research topic.

The most widespread detection framework consists of several stages [10]: (1) a *selection of candidates* (image windows) to be classified as containing a pedestrian or not; (2) the *classification* of such windows; and (3) a *non-maximum suppression* process to remove redundant detections. For videos (4) *tracking* is also used to remove spurious detections, and deriving information like pedestrian motion direction. All these stages are quite relevant and can contribute on their own to achieve a reliable pedestrian detector in terms of processing time and detection performance. However, since the number of candidates per image runs from thousands to hundred of thousands, the classification stage is specially critical in such processing pipeline. Accordingly, most of the work done on image-based pedestrian detection has been focused on classification, *i.e.* given a candidate window decide if it contains a pedestrian or not.

Key components of a pedestrian classifier are the pedestrian descriptors and the machine learning algorithm employed to obtain the classifier. Thus, most works on pedestrian detection have focused on these aspects [5, 10, 3]. The initial task of the learning process consists in collecting examples of pedestrians (*positives*) and background (*negatives*), which is critical since with poor examples even the best combination of descriptors and learning machine cannot provide a good classifier. Accordingly, different pedestrian datasets have been collected through manual annotation (*e.g.* INRIA [2], Daimler [5], Caltech [3], ETH and TUD [18], CVC02 [9]), where *annotating a pedestrian* means to provide its BB at least.

Since manual annotation is a never ending tiresome process, not only for pedestrian detection but for object detection in general, different methods have been proposed to alleviate it. For instance, a new annotation paradigm consists in crowd-sourcing with web-based tools. A well known example is Amazon's Mechanical Turk (MTurk) [13], which allows researchers to define *human intelligence tasks* (HITs:

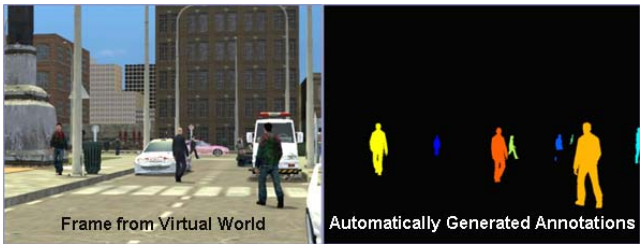

Figure 1. Virtual image with corresponding automatically generated pixel-wise ground truth for pedestrians.

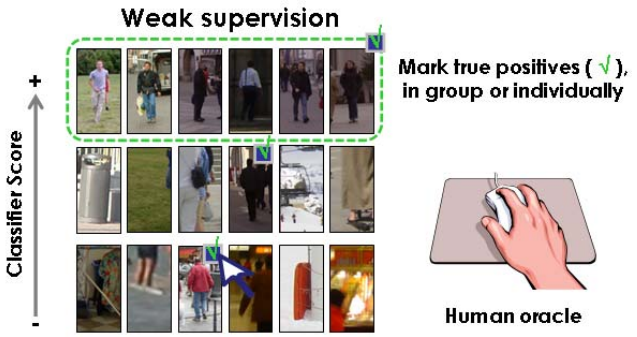

Figure 2. Detections are presented to the oracle ordered by classifier score and in CW size. The oracle marks right detections individually or in groups indicated by initial and final clicks.

what and *how*) of different difficulty (*e.g*. from marking points of interest to drawing polygons) to be taken by human online workers (*turkers*) which are paid for their work. Thousands of annotations can be collected through MTurk. Unfortunately, most turkers are not Computer Vision experts and have not scientific motivation, which makes annotation quality very sensitive to both annotation instructions [4] and economic reguard [15]. In fact, due to such reasons and even to malicious workers, it is necessary to collect multiple annotations from the same image/object and assess annotation quality from them [15]. Thus, *how to collect data on the internet* is a non-trivial question that opens a new research area [15].

Another interesting alternative was presented in [12] for pedestrian detection. In particular, photo-realistic virtual worlds were proposed for collecting training samples. Following such an approach, detailed ground truth is automatically available for each virtual-world pedestrian, *i.e*. its BB and silhouette (Fig. 1). Pedestrian-free images are automatically generated as well. Yet, the challenge consists in achieving good pedestrian detection performance with real-world images using classifiers learned from such virtual-world samples. Even results are rather satisfactory, this procedure shows the *dataset shift* problem [14] since virtual- and real-world images have inherent differences. Therefore, without designing appropriate *domain adaptation* techniques [17] it can be a loss of performance when training in virtual world and testing in real world, *i.e*. as when training with data from a camera and testing with data from another one [16].

We are interested in minimizing the annotation effort required for developing object detectors in general, and pedestrian detectors in particular. Thus, indeed we think that virtual worlds are an interesting framework to explore. However, rather than devising domain adaptation procedures, we propose to use the virtual-world data for developing a pedestrian classifier to be used for collecting pedestrian detections from real-world images. Then a human oracle validates the detections as right or false. The idea is that at the end of the process we can end up with a large number of real-world pedestrian BBs without manually annotating them, *i.e*. the virtual-world-based pedestrian detector provides BBs for us, while the human oracle just provide easy *yes/no*-feedback to validate such BBs. In this paper we show that accurate BBs can be obtained through this procedure, saving a lot of oracle time. Moreover, the procedure is adaptable to work in crowd-sourcing style but allowing to propose a simpler task less prone to errors.

The rest of the paper is organized as follows. In Sect. 2 we detail our proposal. In Sect. 3 we draw our experimental settings. In Sect. 4 we present and discuss the obtained results. Finally, in Sect. 5 we summarize our conclusions.

2. Weakly supervised training

2.1. Our proposal in a nutshell

In this paper we use a virtual city to automatically collect samples for training a pedestrian classifier. Since such a classifier must operate in real-world images we should either design a domain adaptation procedure using a few real-world annotations, or use the classifier to collect many real-world annotations. Both cases require re-training. In this paper, we follow the second approach proposing a *weakly supervised* annotation procedure, *i.e*. pedestrian BBs are not manually annotated. We first train a pedestrian classifier using only virtual-world data. Then, such a classifier collects pedestrian examples from real-world images by detection. A human oracle rejects false detections through an efficient procedure. Thus, at the end of the process we obtain pedestrian examples without requiring manual annotation of BBs. Real-world examples are then used to train the final pedestrian classifier.

In order to learn pedestrian classifiers we employ the components proposed in [2], *i.e*. histograms of oriented gradients (HOG) as descriptors and linear support vector machines (LinSVMs) as base learners. HOG/LinSVM is still a competitive baseline and a key component of state-of-the-art detectors [3]. Under these settings, we show that

Figure 3. Ground truth (top row) and detections (bottom row). Left block of five pedestrians contains detections with classifier score in $[-1, 0)$, those in mid block are in $[0, 1)$, and those in left block correspond to values ≥ 1. In our current settings, left and mid blocks are discarded and only the detections of the right block would arrive to the human oracle for validation (along with some hard negatives).

our weakly supervised approach provides classifiers analogous to their counterparts trained with examples collected by manually annotating the BBs of all the available pedestrians.

2.2. Weakly annotation of BBs

Assume the following definitions of *training sets*:

- *Source domain*. Let $\Im_{\mathcal{V}}^{tr+}$ be the set of available virtual-world images with automatically annotated pedestrians, and $\Im_{\mathcal{V}}^{tr-}$ the set of pedestrian-free virtual-world images automatically generated as well.

- *Target domain*. Let $\Im_{\mathcal{R}}^{tr+}$ be a set of real-world images with non-annotated pedestrians, and $\Im_{\mathcal{R}}^{tr-}$ a set of pedestrian-free real-world images.

Define:

- *Classifier basics*, *i.e*. pedestrian description process ($\mathbf{D}$, *i.e*. features computation) and base learner ($\mathcal{L}$).

- *Detections*, *i.e*. provide a threshold Thr such that an image window is said to contain a pedestrian if its classification score is greater than Thr.

Our weakly supervised training consists of the following steps:

(s1) *Train in virtual world* using $\mathbf{D}$ and $\mathcal{L}$ with samples from $\{\Im_{\mathcal{V}}^{tr+}, \Im_{\mathcal{V}}^{tr-}\}$. Let us term as $\mathcal{C}_{\mathcal{V}}$ the learned classifier and as $\mathcal{D}_{\mathcal{V}}$ its associated detector. Let $\mathcal{T}_{\mathcal{V}}^{tr+}$ be the set of pedestrian examples used for obtaining $\mathcal{C}_{\mathcal{V}}$ (*i.e*. coming from $\Im_{\mathcal{V}}^{tr+}$), and $\mathcal{T}_{\mathcal{V}}^{tr-}$ the set of background examples (*i.e*. coming from $\Im_{\mathcal{V}}^{tr-}$). Examples in $\mathcal{T}_{\mathcal{V}}^{tr+}$ and $\mathcal{T}_{\mathcal{V}}^{tr-}$ are assumed to follow standard steps in the training of pedestrian classifiers, namely, they are in canonical window (CW) size, $\mathcal{T}_{\mathcal{V}}^{tr+}$ includes mirroring, and $\mathcal{T}_{\mathcal{V}}^{tr-}$ includes bootstrapped hard negatives (previous to bootstrapping, the initial classifier is trained with the same number of positive and negative samples). Let $\mathcal{C}$ denote the current classifier during our learning procedure, and $\mathcal{D}$ its associate detector. Now we provide the initialization $\mathcal{C} \leftarrow \mathcal{C}_{\mathcal{V}}$ (thus, $\mathcal{D}$ is $\mathcal{D}_{\mathcal{V}}$ at the start).

(s2) *Weakly annotating real world*. Run $\mathcal{D}$ on $\Im_{\mathcal{R}}^{tr+}$. Show the detections to the human oracle ($\mathcal{O}$) ordered by $\mathcal{C}$ score, and let $\mathcal{O}$ to mark the true detections in groups or individually (Fig. 2), *i.e*. like when selecting visual items with a graphical interface. Equivalently, we could mark false detections, however, usually true detections are quite far less than false ones. We term as $\mathcal{T}_{\mathcal{R}}^{tr+}$ the set of such new pedestrian examples in CW size and augmented by mirroring. Note that we do not annotate BBs here. This means also that miss detections are not provided by $\mathcal{O}$. In order to collect hard false negatives we can just take the false detections in $\Im_{\mathcal{R}}^{tr+}$ (the detections not marked by $\mathcal{O}$). However, for an easier comparison of our proposal with the standard learning methods used in pedestrian detection, we run $\mathcal{D}$ on $\Im_{\mathcal{R}}^{tr-}$ in order to collect real-world negative samples. Let us term such set of samples as $\mathcal{T}_{\mathcal{R}}^{tr-}$. Moreover, by doing so it is not necessary to mark all true positives, since not marked detections are not assumed to be false positives.

(s3) *Retrain in real world*. Train a new classifier $\mathcal{C}$ with the pedestrian examples collected as validated detections, using $\mathbf{D}$ and $\mathcal{L}$. The new pedestrian detector $\mathcal{D}$ is now based on the new $\mathcal{C}$.

During step **s2**, $\mathcal{D}$ is applied for all images in $\Im_{\mathcal{R}}^{tr+}$ and $\Im_{\mathcal{R}}^{tr-}$, then, step **s3** is applied once. During **s2** we take one negative example per each positive one (same cardinality of $\mathcal{T}_{\mathcal{R}}^{tr+}$, and $\mathcal{T}_{\mathcal{R}}^{tr-}$) and leave for step **s3** collecting more hard negatives by training with bootstrapping using the $\Im_{\mathcal{R}}^{tr-}$ pool.

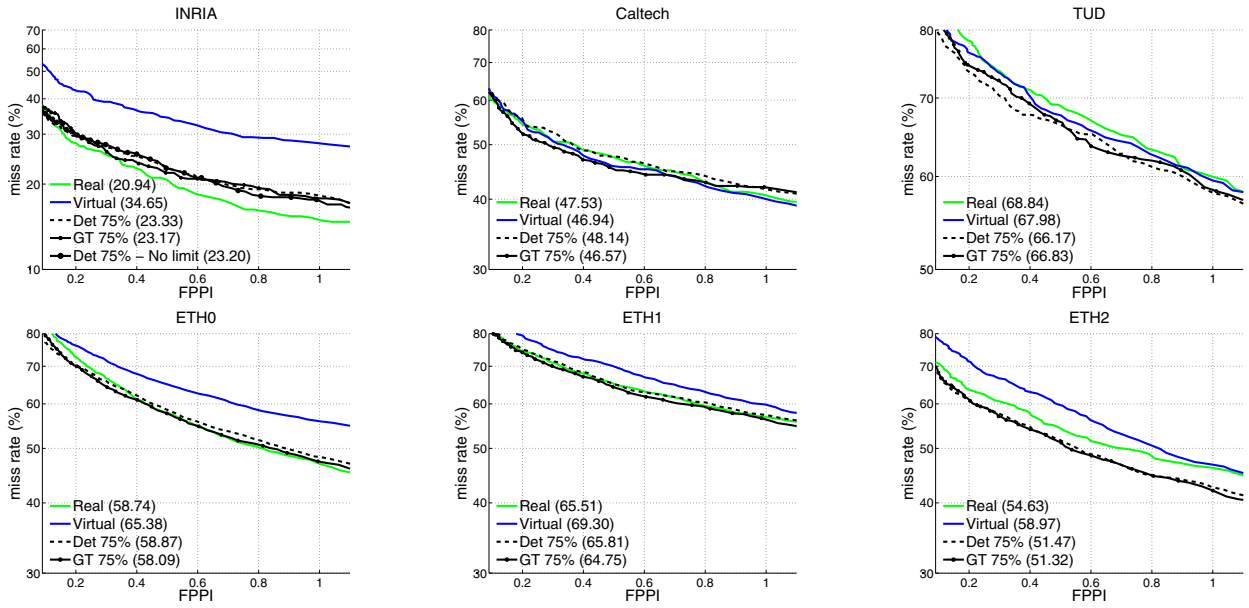

Figure 4. Miss rate vs FPPI curves with the maximum amount of validated detections ('Det' 75%) and corresponding manually annotated BBs ('GT' 75%) from INRIA training set, for different testing sets. 'Real' stands for the case of training with the full INRIA training set, while 'Virtual' refers to training with the virtual-world data. The number in parenthesis is the average miss rate (%) of the respective curve.

3. Experimental settings

3.1. Datasets

For evaluating our weakly supervised annotation proposal, we use the INRIA training set [1] as training real-world dataset, since it is the most well-known baseline and still used to train different state-of-the-art detectors [3]. It contains color images of different resolution (320×240 pix, 1280×960 pix, etc.) with persons photographed in different scenarios (urban, nature, indoor). INRIA data includes a set of training images with the BB annotation of 1208 persons (that can be vertically mirrored to obtain 2416 positive samples). In addition, 1218 person-free images are provided for training. It is worth to note that the BB annotations of the INRIA training and testing sets are considered as precise [15].

As in [12], the virtual-world dataset used to train the corresponding classifier has been generated with Half Life 2 videogame by city driving. It is composed of color images of 640×480 pix. From the provided virtual-world data we mimic the settings of the INRIA training set. Thus, we use 1,208 virtual-world pedestrians that are vertically mirrored to obtain 2416 ones, as well as 1218 pedestrian-free virtual-world images. Of course, since such virtual-world pedestrians have pixel-wise groundtruth, their respective BBs are automatically and accurately computed.

For testing, in addition to INRIA testing set, we use a group of well established pedestrian testing video sequences: Caltech-Testing (Reasonable set) [3], ETH-0,1,2 and TUD-Brussels [18]. So in total, six testing sets.

3.2. Simulating weak annotations

In order to perform fair performance comparison among pedestrian classifiers, for any training we need to rely on the same imaged pedestrians. Thus, we only consider those detections whose BB actually overlap with some corresponding INRIA training ground truth BB (manually annotated). We use the usual PASCAL VOC criterion [6], which defines a level of overlapping of 50%. Paired results termed as 'Det' and 'GT' correspond to pedestrian detectors whose classifiers have been trained with the same pedestrians (INRIA training set), but in the first case the BBs of the pedestrians are given by the validated detections ('Det') while in the second one such BBs are given by the human oracle ('GT').

For the experiments presented in Sect. 4, we also simulate the interaction of the human oracle. In particular, instead of having a person marking the true positives, these are automatically indicated to our system thanks to the INRIA training ground truth. This allows to boost the testing of different alternatives at the current stage of our research. However, in Sect. 4 we evaluate the annotation cost of our proposal by performing experiments with an actual human oracle in the loop.

3.3. Pedestrian classifier and detector

Pedestrian classifiers process image windows, and pedestrian detectors process full images. As we have men-

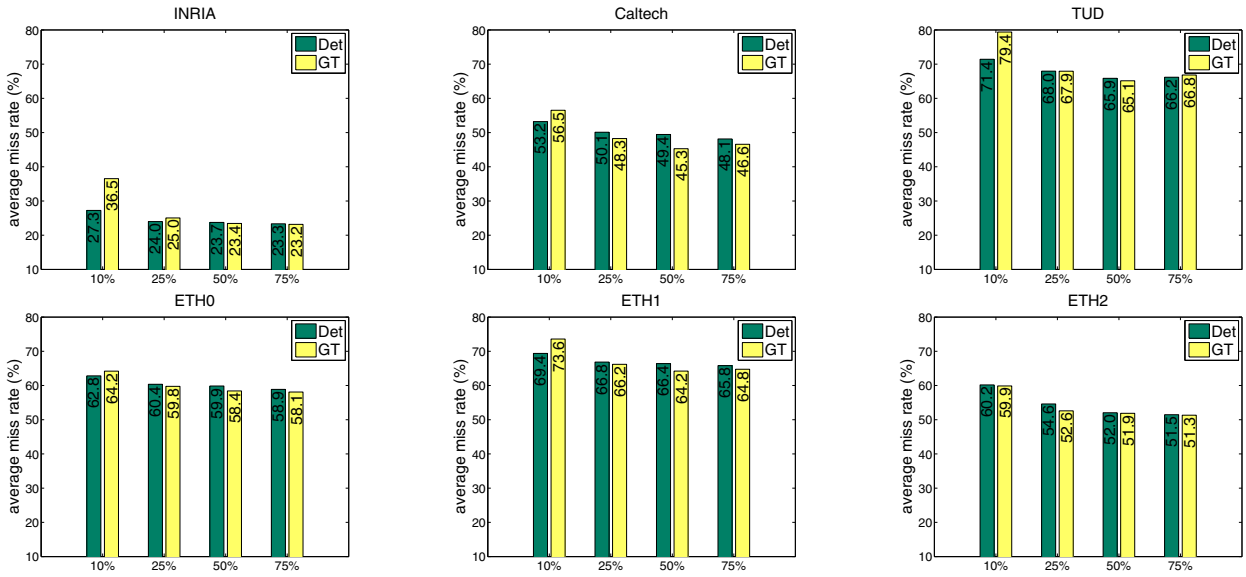

Figure 5. Average miss rate at different testing sets when training with different amounts of validated detections ('Det') and corresponding manually annotated groundtruth ('GT') from INRIA training set. Each bar shows its average miss rate (%).

tioned previously, we rely on HOG descriptors and LinSVM to learn our pedestrian classifiers. For that, we use the same parameters of the original proposal [2]. In order to build the pedestrian detector we also apply the pyramidal sliding window approach as proposed in [1]. However, as done in [8], for all down-sampling operations we rely on standard bilinear interpolation with anti-aliasing, which usually boosts the performance of HOG/LinSVM. Moreover, while for testing we use the same parameters of the pyramidal sliding window than [1] (*i.e.* scale step 1.2, and strides $\triangle x = \triangle y = 8$), for collecting more precise detections to be used in training we follow a finer search (scale step 1.05, and $\triangle x = \triangle y = 4$).

Detection over multiple scales and different positions usually yields several detections which frequently refer to a single object. In order to obtain a unique detection per object (pedestrian), we apply the non-maximum-suppression approach proposed in [11]. Note that we decide if an image window is a detection or not according to the classification score and threshold Thr (Sect. 2.2). Here we have set $Thr = -1.0$, *i.e.* the oracle gives *yes/no*-feedback for windows with classification score ≥ -1.0. Note that for SVM classifiers this is in the ambiguity region. Thus, in practice most of the windows presented to the human oracle for *yes/no*-feedback will be pedestrians, but some of them will be hard negatives.

3.4. Evaluation methodology

In order to evaluate the performance of the pedestrian detectors we reproduce the procedure proposed in [3]. This means that we use performance curves of *miss rate vs. false positives per image*. We focus on the range FPPI=10^{-1} to 10^{0} of such curves, where we provide the *average miss rate* by averaging its values taken at steps of 0.01.

4. Results

Figure 3 provides visual insight about BB localization accuracy for the detections of the virtual-world-based pedestrian detector applied to the INRIA training set. Figure 4 plots the results comparing the performance of the pedestrian detectors resulting from manually annotated BBs *vs*. the BBs resulting from our method (*i.e.* using validated detections) for the same pedestrian examples. For sake of completeness, the results of training with both the full INRIA training set and the virtual-world one are plotted as well. Our validated detections reach almost the 80% of the INRIA training pedestrians, so we decided to set 75% as the limit of our method for such a training set. Figure 5 plots the average miss rate of the 'Det' and 'GT' pedestrian detectors according to different amounts of training data used, being 75% the maximum.

From these results we can draw two main conclusions. On the one hand, the 'Det' and 'GT' performances are so close that we think that BBs from validated detections are as accurate as precise pedestrian BB annotations for developing good classifiers. The difference would be even more negligible by using the HOG/Latent-SVM method for learning deformable part-based models [7], since it is able to refine the annotated BBs. On the other hand, the randomly selected 75% of the annotations seems to already convey the same information than the 100% since the two cases give

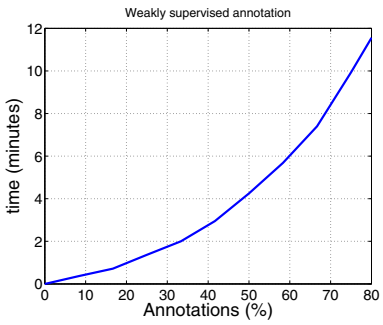

Figure 6. Annotation effort with our weakly supervised method.

rise to a very similar performance.

In order to quantify the human annotation effort of our weakly supervised method, *i.e*. in comparison with the human annotation of BBs, we provide Fig. 6. Note that annotation time is reduced drastically for the human oracle.

For instance, around only 10 minutes are required to annotate the 75% of the pedestrians (906) since no BBs must be provided. We experimented manual annotation of pedestrian accurate BBs and found an average required time of 6 seconds per BB. Thus, annotating the BBs of the 75% would require 90 minutes (9 times more).

5. Conclusions

In this paper we have presented a method for training pedestrian classifiers without manually annotating their required full-body BBs. The two core ingredients are the use of virtual world data, and the design of a weakly supervised procedure to validate detections by just *yes/no* human feedback. Presented results indicate that the obtained classifiers are on pair with the ones based on manual annotation of BBs. Besides, the human intervention is highly reduced in terms of both time and difficulty of the annotation task. Finally, notice also that our method can be applied to other objects.

Acknowledgments

This work is supported by the Spanish MICINN projects TRA2011-29454-C03-01 and TIN2011-29494-C03-02, the Chinese Scholarship Council (CSC) grant No.2011611023 and Sebastian Ramos' FPI Grant BES-2012-058280.

References

[1] N. Dalal. *Finding People in Images and Videos*. PhD thesis, Institut National Polytechnique de Grenoble, 2006. Advisors: Cordelia Schmid and William J. Triggs. 4, 5

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005. 1, 2, 5

[3] P. Dollár, 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*, 34(4):743–761, 2012. 1, 2, 4, 5

[4] I. Endres, A. Farhadi, and D. H. D.A. Forsyth. The benefits and challenges of collecting richer annotations. In *Advancing Computer Vision with Humans in the Loop, CVPR Workshop*, San Francisco, CA, USA, 2010. 2

[5] M. Enzweiler and D.M. Gavrila. Monocular pedestrian detection: survey and experiments. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(12):2179–2195, 2009. 1

[6] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *Int. Journal on Computer Vision*, 88(2):303–338, 2010. 4

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 5

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008. 5

[9] D. Gerónimo, A.D. Sappa, D. Ponsa, and A.M. López. 2D-3D based on-board pedestrian detection system. *Computer Vision and Image Understanding*, 114(5):1239–1258, 2010. 1

[10] D. Gerónimo, A.M. López, A.D. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(7):1239–1258, 2010. 1

[11] I. Laptev. Improving object detection with boosted histograms. *Image and Vision Computing*, 27(5):535–544, 2009. 5

[12] J. Marin, D. Vázquez, D. Gerónimo, and A.M. López. Learning appearance in virtual scenarios for pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010. 2, 4

[13] Amazon Mechanical Turk. www.mturk.com. 1

[14] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence, editors. *Dataset shift in machine learning*. Neural Information Processing. The MIT Press, 2008. 2

[15] T.L Berg, A. Sorokin, G. Wang, D.A. Forsyth, D. Hoeiem, I. Endres, and A. Farhadi. It's all about the data. *Proceedings of the IEEE*, 98(8):1434–1452, 2010. 2, 4

[16] A. Torralba and A.A. Efros. Unbiased look at dataset bias. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011. 2

[17] D. Vázquez, A. López, D. Ponsa, and J. Marín. Cool world: domain adaptation of virtual and real worlds for human detection using active learning. In *Advances in Neural Information Processing Systems – Workshop on Domain Adaptation: Theory and Applications*, Granada, Spain, 2011. 2

[18] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009. 1, 4