Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

A Boosted Multi-Task Model for Pedestrian Detection with Occlusion Handling

Chao Zhu and **Yuxin Peng***
Institute of Computer Science and Technology, Peking University
Beijing 100871, China
{zhuchao, pengyuxin}@pku.edu.cn

Abstract

Pedestrian detection is a challenging problem in computer vision. Especially, a major bottleneck for current state-of-the-art methods is the significant performance decline with increasing occlusion. A common technique for occlusion handling is to train a set of occlusion-specific detectors and merge their results directly. These detectors are trained independently and the relationship among them is ignored. In this paper, we consider pedestrian detection in different occlusion levels as different but related problems, and propose a multi-task model to jointly consider their relatedness and differences. The proposed model adopts multi-task learning algorithm to map pedestrians in different occlusion levels to a common space, where all models corresponding to different occlusion levels are constrained to share a common set of features, and a boosted detector is then constructed to distinguish pedestrians from background. The proposed approach is evaluated on the challenging Caltech pedestrian detection benchmark, and achieves state-of-the-art results on different occlusion-specific test sets.

Introduction

Pedestrian detection is a challenging problem in computer vision, and has attracted a lot of attention for decades, since reliable detection of pedestrians is important in practical applications such as video surveillance, driving assistance and robot navigation. Thanks to more powerful features, more sophisticated models and more effective detection strategies, pedestrian detection has achieved impressive progress in recent years (Dollár et al. 2012). However, occlusion is still a major obstacle for satisfactory detection. Most state-of-the-art methods assume that pedestrians are fully visible or little occluded, and their performances decline significantly with increasing occlusion level. For example, the mean miss rate of the best detector nowadays achieves 36% for pedestrians with no occlusion on the Caltech pedestrian detection benchmark (Dollár et al. 2012), while increasing significantly to 49% for pedestrians with no more than 35% occluded, and further increasing drastically to 79% for pedestrians with 35%-80% occluded. Occlusions occur frequently in real world scenes such as in street scenes or crowded places. For example, according to the Caltech pedestrian benchmark, over 70% of its pedestrians appear occluded in at least one frame of a video sequence and 19% are occluded in all frames, where the occlusion was categorized as heavy (35%-80% occluded) in nearly half of these cases. Therefore, a pedestrian detector capable of handling occlusions is preferred to yield more robust detection results.

One kind of common method for occlusion handling is to estimate the visibility of body parts of pedestrians. This is particularly effective for those approaches based on deformable part-based models (DPM) (Felzenszwalb et al. 2010). DPM sums the scores of all the part detectors within a given window, and identifies the window as positive pedestrian if the summed score is higher than a pre-defined threshold. In the case of occlusion happens, some parts of pedestrians are occluded, resulting in very low scores for the corresponding part detectors, and consequently the summed score will also be low, leading to false negatives. Visibility estimation of parts could help the detection of occluded pedestrians, since it identifies which body parts are visible and which body parts are occluded, where the effects of occluded parts are reduced in the summed score. Usually additional information is required for visibility estimation, *e.g.*, motion information, depth information or segmentation results (Enzweiler et al. 2010). However, such information is not always available.

Another kind of common method for occluded pedestrian detection is to train a set of occlusion-specific detectors, one for each occlusion level. At test time, if the occlusion level is known (a priori or estimated), a corresponding occlusion-specific detector will be applied to detect occluded pedestrians; otherwise, all the occlusion-specific detectors will be applied, and their detection results will then be merged directly. The disadvantages of this method lie in two aspects: on one hand, training such a number of detectors is costly, and the efficiency requirement will be critical as the scale of data or number of classes grows; on the other hand, these occlusion-specific detectors are trained independently, while the relationship among them is ignored.

We believe that the relationship among different occlusion levels should be explored for robust occluded pedestrian detection. For example, heavy occluded samples suffer a serious missing of useful pedestrian information, and the

*Corresponding author.

noisy information extracted from the occluded regions may mislead the detector in the training phase. The information contained in less occluded samples can help regularize it. To this end, we consider pedestrian detection in different occlusion levels as different but related problems, and propose a multi-task model to jointly consider their relatedness and differences. Particularly, we extend the Aggregated Channel Features (ACF) detector (Dollár et al. 2014) to Multi-Task ACF (MT-ACF), which adopts multi-task learning algorithm to map pedestrians in different occlusion levels to a common space, where all models corresponding to different occlusion levels are constrained to share a common set of features, and a boosted detector is then constructed to distinguish pedestrians from background.

To evaluate the proposed approach, we carry out experiments on the challenging Caltech pedestrian detection benchmark (Dollár et al. 2012), and achieve state-of-the-art performances on the most popular "Reasonable" and three occlusion-specific test sets.

Related Work

The great progress has been achieved for pedestrian detection over the last years by the application of different classification approaches, more powerful features and more sophisticated pedestrian models. Nevertheless, limited attention has been paid to address the issue of occlusion handling in the literature. Generally, they can be categorized into two kinds: visibility estimation based and occlusion-specific classifier based.

In order to estimate the visibility of pedestrian parts, various approaches have been proposed (Wu and Nevatia 2005; Leibe, Seemann, and Schiele 2005; Wang, Han, and Yan 2009; Enzweiler et al. 2010; Gao, Packer, and Koller 2011; Ouyang and Wang 2012; 2013; Ouyang, Zeng, and Wang 2013). Some of them adopt detection scores of blocks or parts as input for visibility estimation. In (Wang, Han, and Yan 2009), the authors use a property of HOG features on the human class to infer occluded pixels in a detection window. In (Ouyang and Wang 2012), the authors use a deformable part-based model to obtain the scores of part detectors and the visibilities of parts are modeled as hidden variables. In (Gao, Packer, and Koller 2011), the authors propose a set of binary variables each of which corresponds to a cell indicating the visibility of the cell, *i.e.*, whether the pixels in the cell belong to the object. Some utilize other cues such as segmentation results (Leibe, Seemann, and Schiele 2005; Enzweiler et al. 2010) or depth information (Enzweiler et al. 2010) for visibility estimation. Some apply the recently developed deep learning for visibility estimation. In (Ouyang and Wang 2012), instead of treating visibilities of parts independently, a discriminative deep model is used for learning the visibility relationship among overlapping parts at multiple layers. In (Ouyang, Zeng, and Wang 2013), the authors focus on the situation when several pedestrians overlap in images and occlude each other, and propose a mutual visibility deep model that jointly estimates the visibility status of overlapping pedestrians. However, most of these approaches deal with occlusion handling separately from other components like feature extraction and deformation modeling. In (Ouyang and Wang 2013), the authors explore the interactions among these components, and propose a joint deep learning framework to maximize their strengths.

Other researchers have proposed approaches for handling occlusions by using occlusion-specific classifiers (Wojek et al. 2011; Tang, Andriluka, and Schiele 2012; Mathias et al. 2013). The basic idea is that when applying a fully visible classifier on an occluded detection window, the features extracted from the occluded regions may be miss-leading. In contrast, when training specific classifiers for different occlusion levels, the feature extraction can focus only on the visible regions, thus resulting in improved detection performance. In (Wojek et al. 2011), the authors propose to train a set of classifiers, each one for a specific occlusion. At test time, occlusions are first identified by some techniques, *e.g.*, using segmentation results or depth information, and appropriate classifiers are applied. Despite better detection quality for each case, training many classifiers is costly. This issue is addressed in (Mathias et al. 2013). The authors introduce the idea of spatially biasing feature selection during classifier training. Starting from one biased classifier trained for full-body detection, they reuse training time operations to efficiently build a set of occlusion-specific classifiers, and reduce the computation time by one order of magnitude. In (Tang, Andriluka, and Schiele 2012), the authors propose to train specific detectors for pairs of occluding and occluded objects, and obtain good results for pairs of pedestrians. However, all these approaches train occlusion-specific classifiers independently, while the relationship among them is ignored. On the contrary, in this paper, we consider pedestrian detection in different occlusion levels as different but related problems, and successfully obtained better detection quality by jointly considering their relatedness and differences via a boosted multi-task model.

Multi-Task Aggregated Channel Features with Occlusion Handling

Generally there are two intuitive strategies to handle occluded pedestrian detection. One is to combine samples from different occlusion levels to train a single detector, and the other is to train a set of independent detectors for different occlusion levels. They both have disadvantages of their own. The former strategy considers the commonness between different occlusion levels, while their differences are ignored. Samples from different occlusion levels would increase the complexity of detection boundary, which is probably beyond the ability of a single detector. On the contrary, the latter strategy considers pedestrian detection in different occlusion levels as independent problems, and the relationship among them are missing. The noisy information from heavy occluded pedestrians may mislead the learned detector and degrade the final detection performance.

In this section, we present a robust pedestrian detection method with occlusion handling by considering the relationship of samples from different occlusion levels, including their relatedness and differences, which are captured by a multi-task strategy simultaneously. To deal with the dif-

ferences of different occlusion levels, we adopt multi-task learning algorithms to map pedestrians from different occlusion levels to a common subspace, where all models corresponding to different occlusion levels are constrained to share a common set of features. A shared boosted detector is then constructed in the subspace to capture the relatedness.

Particularly, we apply the idea to the popular ACF detector (Dollár et al. 2014) and propose a multi-task extension of ACF. Here we consider two common occlusion levels: partial occlusion (no more than 35% occluded) and heavy occlusion (35%-80% occluded), as advised in (Dollár et al. 2012). The extension of the strategy for more occlusion levels is straightforward.

Aggregated Channel Features (ACF) Detector

We apply Aggregated Channel Features (ACF) detector as a baseline detector, because of its state-of-the-art detection quality and speed. This approach can be seen as a combination of the classic VJ work (Viola and Jones 2004) and HOG work (Dalal and Triggs 2005). Given an input image, several image channels are first computed, and pixels in every block of each channel are then summed. Features are single pixel lookups in the aggregated channels. Totally 10 image channels are used, including 1 normalized gradient magnitude channel, 6 histogram of oriented gradients channels and 3 LUV color channels.

For detector training, boosting is used to train and combine decision trees over these features to distinguish pedestrians from background. Specifically, ACF uses depth-two decision trees as weak classifiers, where the nodes of the trees are decision stumps, defined by a rectangular region in one of the channels described above together with a threshold over the sum of values over the region. AdaBoost (Friedman, Hastie, and Tibshirani 2000) is then applied to train and combine weak classifiers to obtain a strong classifier. At each training iteration, one weak classifier is built, and the trees are built in a greedy fashion, learning one node at a time. For each node in the weak classifier, a set of candidate nodes are built using a predefined set of regions, and the optimal threshold values are searched exhaustively. The node with the minimum classification error is selected. The training starts with a set of random negative samples and then bootstraps twice, adding additional hard negative samples each time.

For pedestrian detection, the learned strong classifier is applied on the test image using a multi-scale sliding window strategy. The default step size used in the detector is 4 pixels and 8 scales per octave. To obtain the final detections, a non-maximal suppression is used.

Multi-Task ACF Detector

Now we present the multi-task extension of ACF detector for occlusion handling. We consider two binary classification tasks: classifying pedestrians in partial occlusion (denoted as T_P), and classifying pedestrians in heavy occlusion (denoted as T_H). The single task T_P is defined as follows (also similarly for T_H). Let X_P be the instance space, D_P be a distribution over X_P, and $f_P : X_P \to Y_P$ be the target function, given a sample $S_P = \{(x, f_P(x)) \mid x \in X_P\}$, the goal is to find an hypothesis function h_P which minimizes the error $Pr_{x\sim D_P}[h_P(x) \neq f_P(x)]$.

By considering two tasks simultaneously, the multi-task learning problem is defined as follows. Let D be a distribution over $X = X_P \cup X_H$, given a sample $S = S_P \cup S_H$, the goal is to find an hypothesis $h : X \to Y_P \times Y_H$ which minimizes the error $Pr_{<x,i>\sim D}[h_i(x) \neq f_i(x)]$, where $h_i(x)$ is the component of $h(x)$ and $i \in \{P, H\}$. The error is a combination of errors for the two original single tasks. In order to solve this multi-task learning problem by boosting algorithms similar to ACF, multi-task weak classifiers must be built at first.

Multi-Task Weak Classifiers Recall that in ACF detector, depth-two decision trees are used as weak classifiers, where the nodes of the trees are decision stumps, which are defined by a root node and two prediction nodes. Obviously this kind of decision stumps is only suitable for single classification task. Here we follow the idea in (Faddoul et al. 2010) to extend the decision stumps to multi-task forms, which are capable of capturing the relatedness of different tasks.

A multi-task decision stump for two tasks has two levels. The first level is the root node, which is a traditional decision stump for one of the two tasks, *i.e.*, either T_P or T_H. The second level are the two prediction nodes of the first level. For each of the two prediction nodes, it is defined by a traditional decision stump for the other task. It is easy to see that a multi-task decision stump for two classification tasks over the label sets Y_P and Y_H defines a function h from X to $Y_P \times Y_H$. Note that the partition of the input space defined by a multi-task decision stump depends on the two classification tasks.

In order to learn the multi-task decision stumps to build the weak classifiers, we adopt a greedy fashion which is similar to the original ACF detector, *i.e.*, we exhaustively search all possible multi-task decision stumps and keep the one with the minimum error. More specifically, we first loop on the two tasks because each of them can be placed as the root node, and then loop on the three traditional decision stumps (one root node and two prediction nodes) defined in a multi-task decision stump, and then loop on all possible features and all possible values for each feature. If we denote the number of features as N and the maximal number of values for each feature as M, the number of the multi-task decision stumps to be searched is at most $2(2NM)^3$. The algorithm for learning the multi-task decision stumps is presented in Algorithm 1.

MT-ACF Model Learning Once multi-task weak classifiers are constructed by the approach presented in the previous section, we can learn the multi-task ACF detector model by a straightforward adaptation of the original Adaboost to multi-task settings (Faddoul et al. 2010). One difference from single task Adaboost is that we consider two samples S_P and S_H for two tasks T_P and T_H simultaneously, therefore we consider distributions over $S = S_P \cup S_H$. Another difference is that the error is calculated in the multi-task settings. At each iteration t, an hypothesis h^t is a multi-task decision stump, which is a function from X to $Y_P \times Y_H$. Recall that the true error of h^t *w.r.t.* target functions f_P and

Algorithm 1 Multi-Task Decision Stumps Learning

Input: the samples S_P and S_H for the two tasks T_P and T_H, the distribution D over $S = S_P \cup S_H$.
Output: the multi-task stump h with the minimum error.

for $i = 1$ to 2 **do**
  set task T_P (if i=1) or T_H (if i=2) as the root node of the multi-task stump;
  for each of the three decision stumps defined in the multi-task stump **do**
    for each possible feature f **do**
      for each possible value θ for feature f **do**
        if $f > \theta$ **then**
          h_j = $positive$;
        else
          h_j = $negative$;
        end if
        calculate the $err(h_j, S, D)$ as in eq.(1);
      end for
    end for
  end for
end for
return multi-task stump $h = \text{argmin}_{h_j} err(h_j, S, D)$.

Algorithm 2 Multi-Task ACF Model Learning

Input: the number of boosting iterations T, samples for the two tasks S_P and S_H, the label of an example y, the label *w.r.t.* task i for $e \in S_i$ $pred$, a function err that calculates the error of a multi-task stump.
Output: the strong classifier $H(x)$ for MT-ACF detector.

Initialization: initialize the distribution D over $S = S_P \cup S_H$ as $D^1 = init(S)$;
for $t = 1$ to T **do**
  train the multi-task stump h^t by Algorithm 1;
  calculate the error $\epsilon^t = err(h^t, S, D^t)$ as in eq.(1);
  calculate hypothesis weight $\alpha^t = \frac{1}{2}\ln(\frac{1-\epsilon^t}{\epsilon^t})$;
  for $e \in S$ **do**
    update the distribution:
    if $pred(h^t, e) = y(e)$ **then**
      $D^{t+1}(e) = \frac{D^t(e)\cdot\exp(-\alpha^t)}{Z^t}$;
    else
      $D^{t+1}(e) = \frac{D^t(e)\cdot\exp(+\alpha^t)}{Z^t}$;
    end if
  end for
end for
return strong classifier for MT-ACF detector $H(x) = \text{argmax}_{(y_P, y_H)} \sum_{t=1}^{T} \alpha^t h^t(x)$.

f_H is defined as $Pr_{<x,i>\sim D}[h_i^t(x) \neq f_i(x)]$, where $h_i^t(x)$ is the component of $h^t(x)$, $i \in \{P, H\}$. Now, given a distribution D^t over $S = S_P \cup S_H$, the empirical error is defined as:

$$err(h^t, S, D^t) = \sum_{e \in S: pred(h^t, e) \neq y(e)} D^t(e) \qquad (1)$$

where $pred(h^t, e) = h_i^t(e)$ if $e \in S_i$ and $y(e)$ is the label of e in S_i. The empirical error $\epsilon^t = err(h^t, S, D^t)$ is the sum of the weighted error of the hypothesis h^t on the two sets S_P and S_H *w.r.t.* D^t. The algorithm for learning the multi-task ACF model is presented in Algorithm 2.

Experimental Evaluation

The experiments are conducted on the Caltech pedestrian detection benchmark (Dollár et al. 2012), which is by far the largest, most realistic and challenging pedestrian dataset. It consists of approximately 10 hours of 640×480 30Hz video taken from a vehicle driving through regular traffic in an urban environment. The data were captured over 11 sessions, and are roughly divided in half for training and testing. It contains a vast number of pedestrians – about 250,000 frames in 137 approximately minute long segments with a total of 350,000 bounding boxes and 2300 unique pedestrians were annotated. Evaluation is performed on every $30th$ frame. This dataset is challenging for several reasons: on one hand it contains many small pedestrians and has realistic occlusion frequency; on the other hand the image quality is lacking, including blur as well as visible JPEG artifacts (blocks, ringing, quantization) which hurt the accuracy of feature extraction.

Experimental Setup

We follow a common training-testing protocol as in the literature: the pedestrian detector is trained on its training set (set00-set05), and the results are reported on its test set (set06-set10). To train the detector, we choose the image regions labeled as "persons" that are greater than 50 pixels high with different occlusion levels as positive samples, and negative samples are chosen at random locations and sizes from the training images without pedestrians.

For evaluation of the results, we use the bounding boxes labels and the evaluation software (version 3.2.0) provided by Dollár *et al.* on the website[1]. The per-image evaluation methodology is adopted, i.e. all the detection results are compared using miss rate vs. False-Positive-Per-Image (FPPI) curves. The *log-average miss rate* is also used to summarize the detection performance, and is computed by averaging the miss rate at nine FPPI points[2] that are evenly spaced in the log-space in the range from 10^{-2} to 10^{0}. There exist various experimental settings on this dataset to compare detectors in different conditions. In order to validate the effectiveness of the proposed approach, the following experiments will be conducted on the most popular "reasonable" subset (pedestrians of $\geq$ 50 pixels high, fully visible or less than 35% occluded) and three different "occlusion" subsets: "none occlusion" (pedestrians of $\geq$ 50 pixels high, fully visible); "partial occlusion" (pedestrians of $\geq$ 50 pixels high, less than 35% occluded) and "heavy occlusion" (pedestrians of $\geq$ 50 pixels high, 35%-80% occluded).

The training parameters in the proposed approach are set as follows. 2048 weak classifiers are trained and combined to a strong classifier, and the nodes of the decision trees are constructed using a pool of 30,000 candidate regions from

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

[2] The mean miss rate at 0.0100, 0.0178, 0.0316, 0.0562, 0.1000, 0.1778, 0.3162, 0.5623 and 1.0000 FPPI.

Table 1: Log-average miss rate (%) of popular detection methods with occlusion handling on different subsets of Caltech.

	HOG-LBP [ICCV 2009]	**DBN-Isol** [CVPR 2012]	**DBN-Mut** [CVPR 2013]	**Franken** [ICCV 2013]	**JointDeep** [ICCV 2013]	**MT-ACF**
Reasonable	67.77	53.14	48.22	48.68	39.32	**35.81**
None Occlusion	66.29	50.65	45.91	46.29	37.22	**32.44**
Partial Occlusion	79.94	70.71	68.41	67.57	56.82	**50.72**
Heavy Occlusion	96.74	85.98	**80.45**	88.82	81.88	82.37
Mean	77.69	65.12	60.75	62.84	53.81	**50.34**

image samples. The multi-scale models are used to increase scale invariance. Two bootstrapping stages are applied with 5000 additional hard negatives each time.

Comparison with Other Occlusion Handling Strategies

To demonstrate the effectiveness of the proposed multi-task model for occlusion handling, we compare it with other strategies for occluded pedestrian detection on the Caltech benchmark. The compared methods include: (a) detector trained only on partial occluded pedestrians; (b) detector trained only on heavy occluded pedestrians; (c) detector trained on both partial and heavy occluded pedestrians; (d) detector trained on partial and heavy occluded pedestrians independently, and their detection results are then fused.

Fig. 1 shows the detection results on pedestrians of ≥ 50 pixels high, fully visible or no more than 80% occluded. It can be seen that: (1) Heavy occlusion model performs poorly, which is as expected since it is hard to learn an effective pedestrian model due to large occluded portion; (2) Partial occlusion model performs much better than heavy occlusion model, since more useful information of pedestrian is kept for model training; (3) The effect of combining partial and heavy occlusion model depends on the fusion strategy: the improvement is obtained when fusing partial and heavy occlusion data to train a single detector, while no help is provided when training two independent detectors and then fusing their results; (4) The proposed multi-task model outperforms all the other strategies by exploring the relationship of data from different occlusion levels.

Comparison with Popular Detection Methods with Occlusion Handling

We also compare the proposed approach with other popular detection methods with occlusion handling in the literature, including HOG-LBP (Wang, Han, and Yan 2009), DBN-Isol (Ouyang and Wang 2012), DBN-Mut (Ouyang, Zeng, and Wang 2013), Franken (Mathias et al. 2013) and JointDeep (Ouyang and Wang 2013). Table 1 reports the *log-average miss rate* of different detection methods with occlusion handling on "reasonable" and three "occlusion" subsets of the Caltech benchmark. It can be observed that the proposed approach significantly outperforms the other methods on almost all the subsets, except that DBN-Mut and JointDeep are a little better on "heavy occlusion" subset. By averaging the performances on four subsets, the proposed approach outperforms the other methods by at least 3.5%.

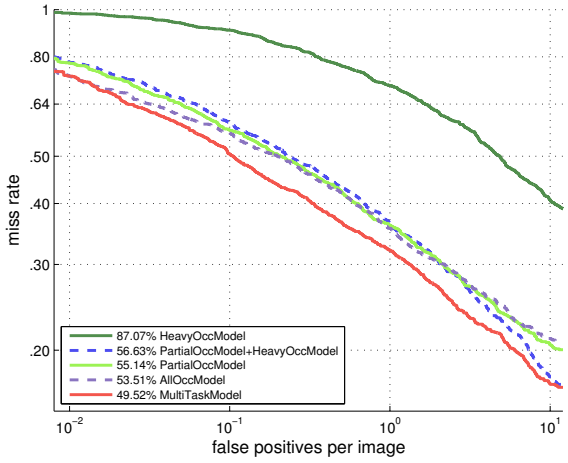
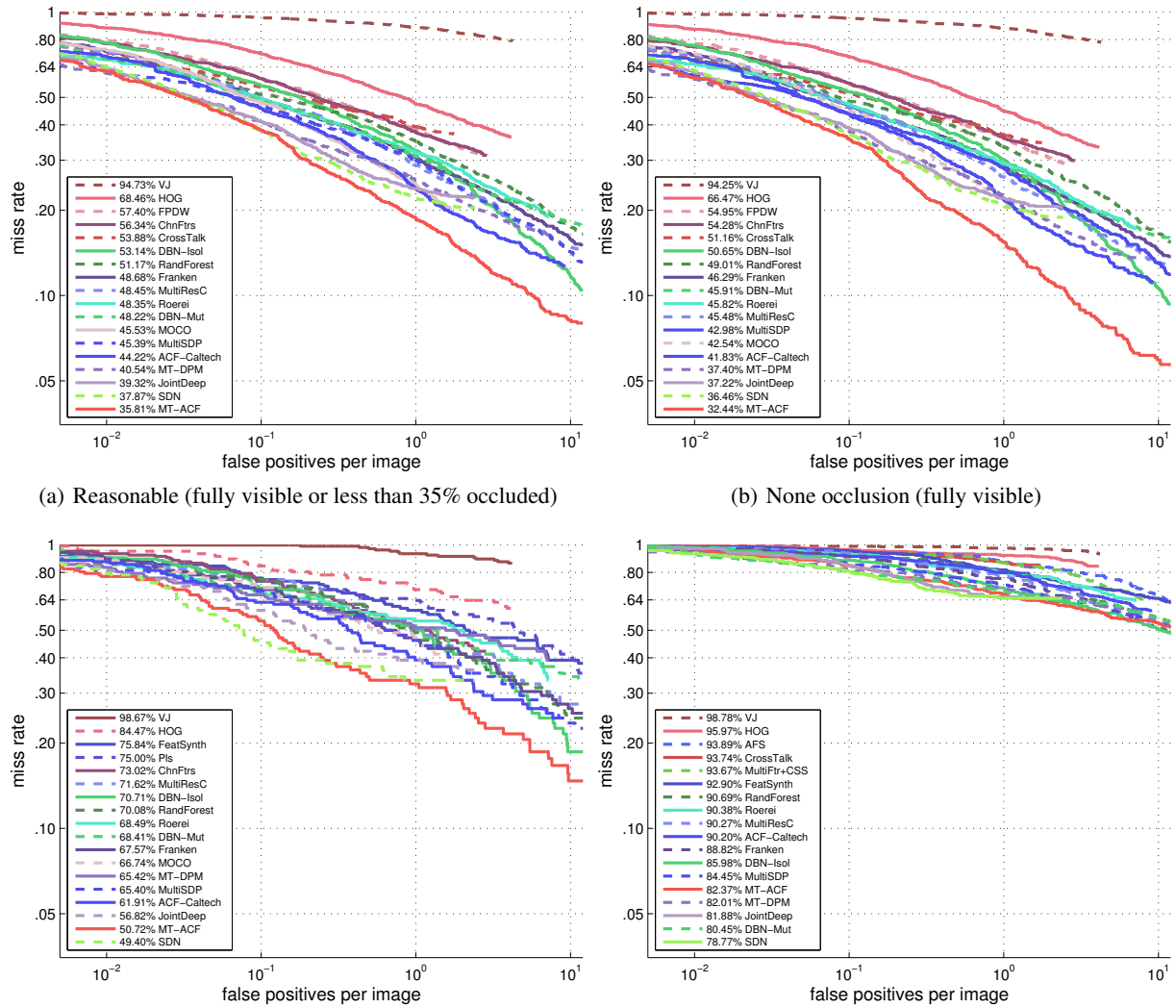


Figure 1: Comparison of different occlusion handling strategies (the numbers in the legend indicate the *log-average miss rate*).

Comparison with State-of-the-art Methods

In this section, we compare the proposed approach with the state-of-the-art detection methods in the literature, including VJ (Viola, Jones, and Snow 2005), HOG (Dalal and Triggs 2005), AFS (Levi, Silberstein, and Bar-Hillel 2013), ChnFtrs (Dollár et al. 2009), ConvNet (Sermanet et al. 2013), CrossTalk (Dollár, Appel, and Kienzle 2012), FeatSynth (Bar-Hillel et al. 2010), FPDW (Dollár, Belongie, and Perona 2010), HikSVM (Maji, Berg, and Malik 2008), LatSVM (Felzenszwalb et al. 2010), MultiFtr (Wojek and Schiele 2008), MOCO (Chen et al. 2013), MT-DPM (Yan et al. 2013), MultiResC (Park, Ramanan, and Fowlkes 2010), pAUCBoost (Paisitkriangkrai, Shen, and van den Hengel 2013), Pls (Schwartz et al. 2009), PoseInv (Lin and Davis 2008), Roerei (Benenson et al. 2013), Shapelet (Sabzmeydani and Mori 2007), RandForest (Marín et al. 2013), MultiSDP (Zeng, Ouyang, and Wang 2013), ACF (Dollár et al. 2014), SDN (Luo et al. 2014), and the methods mentioned in the previous section. Note that for fair comparisons, we focus on the methods which detect pedestrians on static images, and exclude the results of using additional motion or contextual information. We obtain the results of these methods from the same website as the evaluation software.

Fig. 2 presents the ROC curves (miss rate vs. FPPI) and the corresponding *log-average miss rate* (reported in the legend of the figure) of different methods on four subsets of



(a) Reasonable (fully visible or less than 35% occluded)

(b) None occlusion (fully visible)

(c) Partial occlusion (less than 35% occluded)

(d) Heavy occlusion (35%-80% occluded)

Figure 2: Comparison with state-of-the-art methods on the Caltech benchmark.

the Caltech benchmark. Note that only the results of top 16 methods plus the classic VJ and HOG are displayed in the figure due to the space limitation. We can clearly see that: (1) On "reasonable" and "none occlusion" subsets, the proposed approach outperforms all the other state-of-the-art methods both in terms of the ROC curves and the *log-average miss rate* by at least 2.06% and 4.02% respectively; (2) On "partial occlusion" subset, the proposed approach achieves 50.72% in terms of the *log-average miss rate*, which significantly outperforms all the other methods except only SDN. Moreover, if we look at the ROC curves, our approach achieves the best performances (the lowest miss rates) among all the methods when FPPI ≥ 0.32; (3) On "heavy occlusion" subset, the proposed approach achieves 82.37% in terms of the *log-average miss rate*, which performs better than most of the other methods except MT-DPM, JointDeep, DBN-Mut and SDN. These approaches are also optimized for occlusion handling either by discriminative deep models (JointDeep, DBN-Mut and SDN) or by deformable part models (MT-DPM) which are more flexible and robust against occlusion benefiting from the deformable parts.

Conclusions

In this paper, we consider pedestrian detection in different occlusion levels as different but related problems, and propose a multi-task model to jointly consider their relatedness and differences. The proposed model adopts multi-task learning algorithm to map pedestrians in different occlusion levels to a common space, where all models corresponding to different occlusion levels are constrained to share a common set of features, and a boosted detector is then constructed to distinguish pedestrians from background. The proposed approach is evaluated on the challenging Caltech pedestrian detection benchmark, and achieves state-of-the-art results on different occlusion-specific test sets.

Acknowledgments

This work was supported by National Hi-Tech Research and Development Program (863 Program) of China under Grants 2014AA015102 and 2012AA012503, National Natural Science Foundation of China under Grant 61371128, Ph.D. Programs Foundation of Ministry of Education of China under Grant 20120001110097, and China Postdoctoral Science Foundation under Grant 2014M550560.

References

Bar-Hillel, A.; Levi, D.; Krupka, E.; and Goldberg, C. 2010. Part-based feature synthesis for human detection. In *ECCV*, 127–142.

Benenson, R.; Mathias, M.; Tuytelaars, T.; and Gool, L. J. V. 2013. Seeking the strongest rigid detector. In *CVPR*, 3666–3673.

Chen, G.; Ding, Y.; Xiao, J.; and Han, T. X. 2013. Detection evolution with multi-order contextual co-occurrence. In *CVPR*, 1798–1805.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*, 886–893.

Dollár, P.; Appel, R.; and Kienzle, W. 2012. Crosstalk cascades for frame-rate pedestrian detection. In *ECCV*, 645–659.

Dollár, P.; Belongie, S.; and Perona, P. 2010. The fastest pedestrian detector in the west. In *BMVC*, 1–11.

Dollár, P.; Tu, Z.; Perona, P.; and Belongie, S. 2009. Integral channel features. In *BMVC*, 1–11.

Dollár, P.; Wojek, C.; Schiele, B.; and Perona, P. 2012. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* 34(4):743–761.

Dollár, P.; Appel, R.; Belongie, S.; and Perona, P. 2014. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*

Enzweiler, M.; Eigenstetter, A.; Schiele, B.; and Gavrila, D. M. 2010. Multi-cue pedestrian classification with partial occlusion handling. In *CVPR*, 990–997.

Faddoul, J. B.; Chidlovskii, B.; Torre, F.; and Gilleron, R. 2010. Boosting multi-task weak learners with applications to textual and social data. In *ICMLA*, 367–372.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D. A.; and Ramanan, D. 2010. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(9):1627–1645.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2000. Additive logistic regression: a statistical view of boosting. *Annals of Statistics* 95(2):337–407.

Gao, T.; Packer, B.; and Koller, D. 2011. A segmentation-aware object detection model with occlusion handling. In *CVPR*, 1361–1368.

Leibe, B.; Seemann, E.; and Schiele, B. 2005. Pedestrian detection in crowded scenes. In *CVPR*, 878–885.

Levi, D.; Silberstein, S.; and Bar-Hillel, A. 2013. Fast multiple-part based object detection using kd-ferns. In *CVPR*, 947–954.

Lin, Z., and Davis, L. S. 2008. A pose-invariant descriptor for human detection and segmentation. In *ECCV*, 423–436.

Luo, P.; Tian, Y.; Wang, X.; and Tang, X. 2014. Switchable deep network for pedestrian detection. In *CVPR*.

Maji, S.; Berg, A. C.; and Malik, J. 2008. Classification using intersection kernel support vector machines is efficient. In *CVPR*.

Marín, J.; Vázquez, D.; López, A. M.; Amores, J.; and Leibe, B. 2013. Random forests of local experts for pedestrian detection. In *ICCV*, 2592–2599.

Mathias, M.; Benenson, R.; Timofte, R.; and Gool, L. J. V. 2013. Handling occlusions with franken-classifiers. In *ICCV*, 1505–1512.

Ouyang, W., and Wang, X. 2012. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, 3258–3265.

Ouyang, W., and Wang, X. 2013. Joint deep learning for pedestrian detection. In *ICCV*, 2056–2063.

Ouyang, W.; Zeng, X.; and Wang, X. 2013. Modeling mutual visibility relationship in pedestrian detection. In *CVPR*, 3222–3229.

Paisitkriangkrai, S.; Shen, C.; and van den Hengel, A. 2013. Efficient pedestrian detection by directly optimizing the partial area under the roc curve. In *ICCV*, 1057–1064.

Park, D.; Ramanan, D.; and Fowlkes, C. 2010. Multiresolution models for object detection. In *ECCV*, 241–254.

Sabzmeydani, P., and Mori, G. 2007. Detecting pedestrians by learning shapelet features. In *CVPR*.

Schwartz, W. R.; Kembhavi, A.; Harwood, D.; and Davis, L. S. 2009. Human detection using partial least squares analysis. In *ICCV*, 24–31.

Sermanet, P.; Kavukcuoglu, K.; Chintala, S.; and LeCun, Y. 2013. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*, 3626–3633.

Tang, S.; Andriluka, M.; and Schiele, B. 2012. Detection and tracking of occluded people. In *BMVC*, 1–11.

Viola, P. A., and Jones, M. J. 2004. Robust real-time face detection. *International Journal of Computer Vision* 57(2):137–154.

Viola, P. A.; Jones, M. J.; and Snow, D. 2005. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision* 63(2):153–161.

Wang, X.; Han, T. X.; and Yan, S. 2009. An hog-lbp human detector with partial occlusion handling. In *ICCV*, 32–39.

Wojek, C., and Schiele, B. 2008. A performance evaluation of single and multi-feature people detection. In *DAGM-Symposium*, 82–91.

Wojek, C.; Walk, S.; Roth, S.; and Schiele, B. 2011. Monocular 3d scene understanding with explicit occlusion reasoning. In *CVPR*, 1993–2000.

Wu, B., and Nevatia, R. 2005. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, 90–97.

Yan, J.; Zhang, X.; Lei, Z.; Liao, S.; and Li, S. Z. 2013. Robust multi-resolution pedestrian detection in traffic scenes. In *CVPR*, 3033–3040.

Zeng, X.; Ouyang, W.; and Wang, X. 2013. Multi-stage contextual deep learning for pedestrian detection. In *ICCV*, 121–128.