Is Pedestrian Detection Really a Hard Task?[*]

Helmut Grabner Peter M. Roth Horst Bischof
Institute for Computer Graphics and Vision
Graz University of Technology
{hgrabner, pmroth, bischof}@icg.tugraz.at

Abstract

In this paper we present a simple approach for person detection in surveillance for static cameras. The basic idea is to train a separate classifier for each image location which has only to discriminate the object from the background at a specific location. This is a considerably simpler problem than the detection of persons on arbitrary backgrounds. Therefore, we use adaptive classifiers which are trained online. Due to the reduced complexity we can use a simple update strategy that requires only a few positive samples and is stable by design. This is an essential property for real world applications which require operation for 24 hours a day, 7 days a week. We demonstrate and evaluate the method on publicly available sequences and compare it to state-of-the-art methods which reveals that despite the simple strategy the obtained performance is competitive.

1. Introduction

Tracking and detection of persons are important tasks for many computer vision applications ranging from from visual surveillance to video editing. In this paper we focus on the detection task, i.e., finding the location of people in the image. Due to the variability in the appearance of persons (e.g., clothing, pose in the scene, illumination) and in the background (e.g., clutter, occlusions, moving objects) this task is inherently difficult.

Early approaches used change detection (motion detection) to find moving persons. Therefore, a background model was estimated and based on blob analysis moving pixels were grouped into people hypotheses (e.g., [16]). These approaches are only effective if there are no camera movements, the density of the people is low (such that the persons are isolated), and the background variability is not too harsh. Hence, for more complex scenarios motion detection can not be applied.

On the other end of the complexity spectrum we find several approaches that are based on machine learning methods. Starting with the work of Papageorgiou et al. [12] who applied a Support Vector machine (SVM) and Haar wavelet features to train a person detector. Further on, people have used either global features such as Gavrilas edge templates [5], shape features (e.g., Felzenszwalb et al. [3], Leibe et al. [8]), or local approaches such as combining Haar wavelets [18] or Viola et al. who used in addition to static features also local motion features [19]. Mikolajczek et al. [10] and Dalal and Triggs [4] have used histograms of edge orientations. These histogram features have been carefully analyzed. The developed sophisticated histogram of gradient (HOG) features that are very well suited for pedestrian detection. Wu and Nevatia [21] have achieved promising detection results with edgelet features as well as Sabzmeydani et al. by learning shaplet features [15]. Recently, Tuzel et al. [17] used covariance matrices as descriptors. The goal of all of these approaches is to build a generic person detector which should be applicable for different scenarios and tasks. Thus, for these methods a large training set is required that captures all variability of persons and backgrounds. Obviously, the main limitations of these detectors is to gather a representative training set.

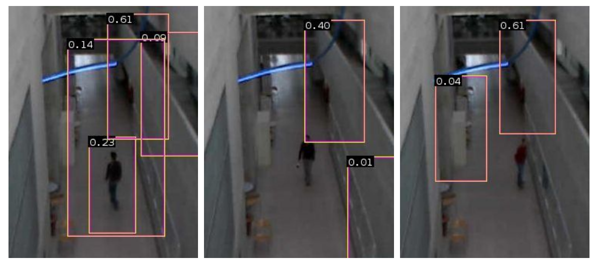


Figure 1: State-of-the-art person detector (Dalal and Triggs [4]) applied on a typical scene. Due to the number of misses and false positives the system can not applied for real world applications.

In this paper we ask the question "Is this task feasible?" and more important "Is this the only way to obtain a reliable person detector?". To motivate these questions let us look at results obtainable by a state-of-the-art detector. Therefore, we applied the detector of Dalal and Triggs [4] that

[*]This work has been supported by the Austrian Joint Research Project Cognitive Vision under projects S9103-N04 and S9104-N04 and FFG project AUTOVISTA 813395 under the FIT-IT program.

can be downloaded from the Internet[1] in a classical not too complex surveillance scenario. Representative results are shown in Figure 1. One can see that a lot of false positives are returned and that persons are missed. Hence, this example shows that there is still a long way to go in order to obtain a generic detector with satisfactory performance.

But do we really need a generic detector for visual surveillance? For a scenario as shown in Figure 1 we have fixed mounted cameras looking always at the same scene which is a considerable simplification for the task of person detection. Hence, we do not need to solve the generic person detection task, we just need to solve it for that particular scene. That is the reason why people have started to look at approaches that can train a person detector for a particular scenario on-line. For more details see Section 2.2. By limiting the detection task to a specific scene the task becomes easier and less training samples are required. On the other hand on-line unsupervised learning methods tend to wrong updates which reduces the performance of the detector. The detector might start to drift and would end in an unreliable state. None of the mentioned approaches reported extended periods (one or more days) of learning. But this is required for realistic surveillance tasks and this is the problem we attack in this paper. We develop an on-line learning method for static surveillance cameras that does not suffer from the drifting problem. In fact, we can guarantee that whether the false positive rate is increased nor that the recall is decreased if the system is running for a longer period of time. Moreover, we show that we only require a very small number of training samples (as few as one). This remarkable results are obtained by further reducing the complexity of the person detection task and by using a fixed (statistically correct) update strategy.

The rest of the paper is organized as follows. In Section 2 we shortly review fixed and adaptive methods for person detection. Section 3 introduces our simplified grid-based person model together with the applied fixed update strategy. Experimental evaluations on different datasets compared to existing approaches are given in Section 4. Finally, we conclude and summarize the paper in Section 5.

2. Scene Adaption Approaches

In this paper we consider only patch-based pedestrian detection. Usually a (discriminative) classifier is build from positive and negative training samples[2]. The positive samples correspond to different appearances of pedestrians whereas the negative samples are usually samples from a very large database of images which does not contain pedestrians at all. During evaluation the classifier is evaluated on all possible locations in the test image (at different scales). Since for a single person various overlapping detections are reported, those are combined in a post-processing step via a simple non-maxima suppression or a mean shift based clustering (e.g., [17]).

Generally, two different types of models can be distinguished: (a) fixed models which are trained off-line and (b) adaptive models which are trained on-line. An illustrative overview of these patch-based detection approaches is given in Figure 2. The dark highlighted grid elements on the left side illustrate the patches where the trained detector should be applicable. On the right side the training datasets (positive and negative samples) of each approach are sketched.

2.1. Fixed Models

Given a fixed training set $\mathcal{X} = \{\langle \mathbf{x}_1, y_1\rangle, ..., \langle \mathbf{x}_L, y_L\rangle \mid \mathbf{x}_i \in \mathbb{R}^m,\ y_i \in \{-1, +1\}\}$ of L samples. A fixed detector is build using an off-line training algorithm. If the detector is trained once (the training time is irrelevant) the parameters of the model (e.g. chosen features, weights,...) are saved and the model can be used for the detection task. Since the parameters are fixed the detector has to handle all possible situations and has to perform well at any time on all possible scenes and all positions in the image. Thus, to finally get a representative model a huge amount of training data is necessary. This principle is illustrated in Figure 2(a).

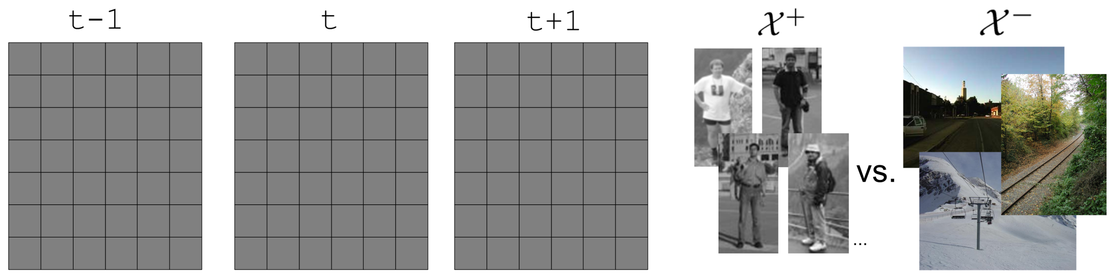
2.2. Adaptive Models

To overcome these problems an adaptive detector using an on-line learning algorithm can be applied. Hence, the system can adapt to changing environments (e.g., changing illumination conditions) and these variations need not to be handled by the model. Compared to a fixed model the detection task is much easier since the detector has "only" to distinguish the positive class (persons) from the background of a specific scene. Thus, the variability of the background as well as the number of required training samples is reduced which is illustrated in Figure 2(b).

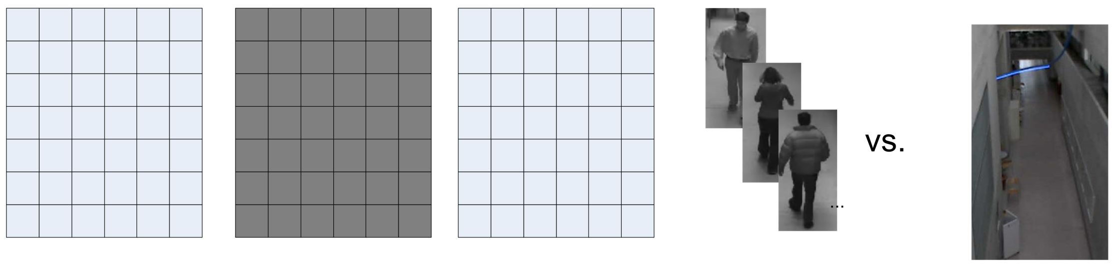
But adaptive systems have one main disadvantage: new unlabeled data has to robustly be included into an already build model. More formally, at time t given a classifier C_{t-1} and an unlabeled example $\mathbf{x}_t \in \mathbb{R}^m$. The classifier predicts a label $y_t \in \{+1, -1\}$ for $\mathbf{x}_t$ which can further be used by an "analyzer" to generate the label $\hat{y}_t$ which is then used to update the classifier: $C_t = \text{update}(C_{t-1}, \langle \mathbf{x}_t, \hat{y}\rangle)$. Different update schemes for on-line learning are illustrated in Figure 3. $\hat{y}$ may be wrong and thus such an update decreases the classification performance. In the following we briefly review common update strategies for learning from unlabeled samples.

[1] `http://pascal.inrialpes.fr/soft/olt`, August 29, 2007. The experiments were performed using the version that was available before August 2007.

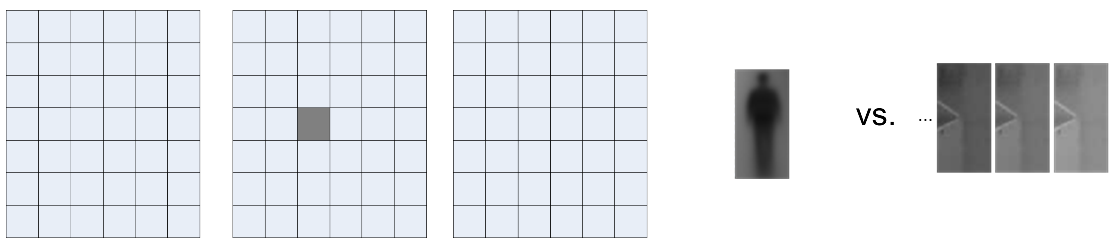
[2] In this paper we only consider the binary classification case.



(a) off-line



(b) on-line



(c) our approach: grid based

Figure 2: Different approaches of a pedestrian detector and the corresponding trainings sets. The gray blocks highlight the regions in both, time and location where the classifier has to perform well: (a) fixed detector, (b) a scene specific on-line trained classifier, and (c) the proposed grid-based detector which specializes for both, time and space.

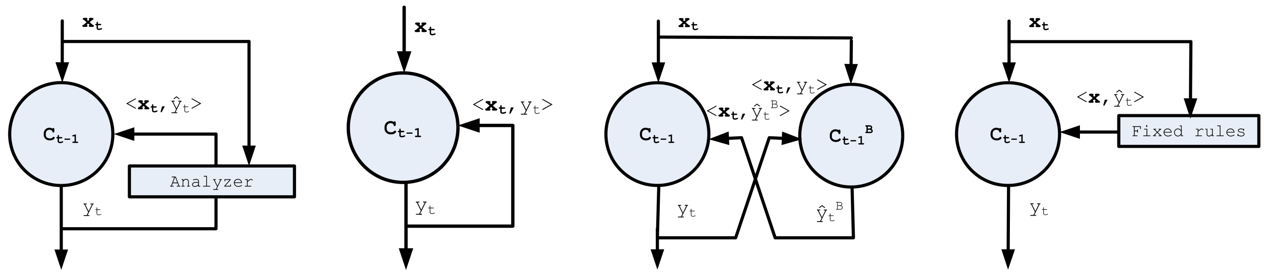
Self-training In a self-training framework the current classifier evaluates an input sample and predicts a label which is then directly used to update the classifier. Hence, the classifier teaches itself by its own predictions. In general, such methods suffer from the drifting problem since there is no re-active process included.

Co-training In a co-training framework (e.g., [2], [9]) two classifiers are trained in parallel using different views of the data. The confident predicted labels are used to update the other classifier, respectively. The main drawback is the assumption that the two classifiers are statistically independent.

Autonomous supervision Autonomous supervision can be considered a simplified variant of co-training. The results obtained by the classifier are verified by a "sophisticated" analyzer and if the obtained labels are confident the samples are used for updating the classifier. Nair and Clark [11] proposed to use motion cues for this purpose. Roth et al. [14] extended this idea and additionally applied a generative model for verification. In contrast, Wu and Nevatia [20] used local parts of the object in order verify the detections and to improve the detection results over time.

3. Grid-based Person Detector

Adaptive approaches such as reviewed in the last section suffer from the drifting problem. A classifier that was trained using many incorrect updates would yield many false positives and/or the detection rate would decrease. Further on, since the classifier responds is used for labeling new samples this would result in a self-fulfilling prophecy. In fact, self-training or co-training which rely on a direct feedback of the current classifier must be avoided. Thus, the main goal of this paper is to define an update strategy that does not suffer from the drifting problem. The key idea is to reduce the complexity of the pedestrian detection prob-



(a) general approach (b) self-learning (c) co-learning (d) our approach: fixed-learning

Figure 3: On-line learning for updating a classifier C: (a) General approach. Depending on the design of the analyzer a wide range of methods can be obtained; from supervised learning (the analyzer is an oracle) to (b) self-learning where the analyzer is hot-wired. (c) Co-learning methods try to handle the drifting problem. Our proposed method (d) does not take into account the classifier response $y_t = C_{t-1}(\mathbf{x}_t)$ for delivering the update. Thus, we have neither direct nor indirect feedback.

lem such that a very simple and fixed update strategy can be applied.

3.1. Reducing the Complexity

A generic person detector must enable the detection of a person in any image at any position (see Figure 2(a)). For adaptive methods the complexity is reduced since the detector is evaluated only within a specific scene (see Figure 2(b)). To further reduce the complexity of the detection problem the detector can be limited to a specific position in the image (see Figure 2(c)). Hence, the task of the detector is to discriminate a person from the background at a specific location in the image. This is a much simpler problem than learning a generic person detector for all situations and it is even much simpler than developing a detector for a specific scene.

For practical realization a fixed highly overlapping grid (both in location and scale) is placed in the image. Each grid element $i = 1, \ldots, N$ corresponds to a classifier C_i which is responsible only for its underlying image patch. Please note, that the speed of evaluation is not significantly decreased. In fact, it does not matter if one classifier is evaluated on all image patches or if each image patch is evaluated using a separate classifier. But since a great number of classifiers has to be stored the memory requirements are increased. Nevertheless, each classifier is simple (has fewer parameters) and can be saved and evaluated more efficiently. The overall system - the classifier grid - is depicted in Figure 4.

3.2. Fixed Update Strategies

Once the classifier grid defined in the previous section was initialized randomly the classifiers can be updated on-line. In the following we discuss the update strategy for a single patch (on classifier C_i). Since updates should be generated

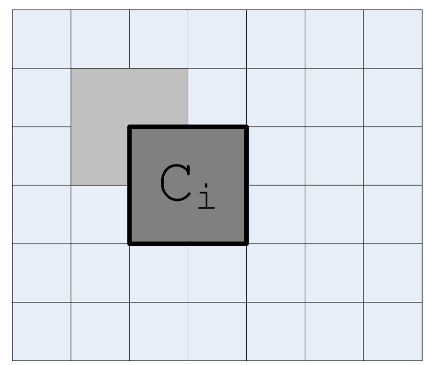


Figure 4: Highly overlapping classifier grid. On each defined image position (location and scale) a classifier is placed which analyzes the underlying image patch.

without a feedback of $C_{i,t-1}$ we make use of the following (very simple) observations:

Positive updates: Given a set of positive (hand) labeled examples $\mathcal{X}^+$. Then, using

$$\langle \mathbf{x}, +1 \rangle, \qquad \mathbf{x} \in \mathcal{X}^+ \tag{1}$$

to update the classifier is a correct positive update. The set can by quite small; in the extremal case (as we will show in the experiments) it contains only one positive sample. The only assumption is that $\mathcal{X}^+$ is a representative set. Roughly speaking, each possible appearance should be captured by this subset.

Negative updates: The probability that a person is present on patch $\mathbf{x}_i$ is given by

$$P(\mathbf{x}_i = \text{person}) = \frac{\#p_i}{\Delta t} \ , \tag{2}$$

where $\#p_i$ is the number of persons entirely present in a particular patch within the time interval Δt. Thus, the negative update with the current patch

$$\langle \mathbf{x}_{i,t}, -1 \rangle \tag{3}$$

is correct most of the time (wrong with probability $P(\mathbf{x}_i = \mathrm{person})$). The probability of a wrong update for this particular image patch is indeed very low.

By using these update rules we avoid the dependencies between the updates and the current model. Since the positive updates are per definition always correct the only remaining problem is that occasionally false negative updates may be carried out. Hence, the applied on-line learning method (a) must cope with some (low) label noise and (b) must have fading memory (forgetting).

3.3. Discussion

The main goal of this paper was to develop an adaptive person detection system that can run over a long time period. Therefore, necessarily a drifting of the underlying classifier must be prevented. This is achieved by using fixed update rules. Since only a small number of positive samples is available the variability of persons can not be modeled well and the results would not be convincing. But the complexity of the detection problem can be reduced such that a single detector has only to distinguish between a single background patch and a person. By combining both approaches we finally get a person detection system that is stable even when running over a long period of time. In fact, we represent the background with high accuracy for each specific grid element and the positive samples provide a suitable “threshold” for the decision.

A different view of this approach can be obtained when discussing background models. Recently, we proposed a classifier-based background model [7]. Therefore, we subdivide the input images into small overlapping blocks and compute a discriminative classifier for each block. The background distribution is estimated via examples taken from the patches whereas the foreground distribution was analytically pre-calculated. In this way non-background blocks, i.e., blocks representing a foreground object, can be identified.

4. Experiments

In the following we will demonstrate the benefits of the presented approach compared to existing methods. Therefore, we split the experiments into two main parts. First, we give a detailed evaluation of the method on a simple but representative test scene. Second, to show that we obtain state-of-the-art detection rates we have applied the proposed method on publicly available benchmark data sets and compared the results to other available person detectors.

4.1. Description of Experiments

First, since the approximative size of the persons in the scene is needed we manually estimated the ground-plane for our each scenes. Of course the ground plane can be estimated automatically (e.g., [13]) as well. Based on these estimation a grid of detectors using an overlap-rate of 90% is initialized. In particular, we use On-line Boosting for Feature Selection [6] to compute the classifiers for these detectors; but any other on-line learning algorithm (e.g., Winnow) can be applied. To compute the grid-based classifier we use only 10 selectors each using a set of 20 weak classifiers. Haar-like features are used because they can be evaluated very fast using the integral data structure. The thus obtained grid of detectors is evaluated and updated whenever a new frame arises. The set of positive samples was reduced to a single image that was obtained by averaging of approximative 100 images of persons.

For the experiments the system described above was evaluated on three different test data sets. In addition, two other person detectors were run the same test sequences, i.e., the Dalal and Triggs person detector[1] [4] and a person detector trained using Conservative Learning [14]. The Dalal and Triggs detector is a generic detector that does not use any previous knowledge. In contrast, the Conservative Learning detector uses scene information to adapt the to a specific scene. Hence, it is less generic but the performance for the particular scene is increased. To avoid multiple detections non-maxima suppression was applied.

For the Dalal and Triggs detector the original parameters[3] are used. To allow a fair comparison in a post-processing step all detections were removed that do not fit to the estimated ground-plane. In fact, a detection was removed if the scale was smaller than 50% or greater than 150% of the expected patch-size. Please note, this post-processing does not reduce the recall since these detections would be counted as false positives. To speed up the computation for Conservative Learning the estimated ground-plane is used to define the patches for the sliding window approach.

For a quantitative evaluation, we use recall-precision curves (RPC) [1]. Therefore, the number of true positives TP and the number of false positives FP are computed based on the given ground-truth. A detection is accepted as true positive if it fulfills the overlap as well as the relative distance criterion where for both criteria the parameters (minimal overlap, maximal relative distance) are set to 50%. The precision rate PR describing the accuracy of the detections is calculated by

$$PR = \frac{TP}{TP + FP} \tag{4}$$

[3]For the computation of precision-recall-curves the parameter *ExtraOption2* was changed to “$-t\ -2\ -m\ 0$” to get a higher recall.

whereas the recall rate RR describing the number of positive samples that were correctly classified is given by

$$RR = \frac{TP}{detections} , \tag{5}$$

where $detections$ is total number of persons in the ground-truth. Finally, to evaluate the detection results we plot the recall rate RR against $1 - PR$.

4.2. Toy Example

First, to discuss the fundamental properties we demonstrate the proposed method on a simple but still representative dataset. Therefore, we have created a sequence of 325 frames showing a corridor in a public building and an according ground-truth (a typical frame was shown in Figure 1). The scene is "simple" since at most one person is present per frame. Hence, there are no occlusions or partial detections that may decrease the recall or the precision.

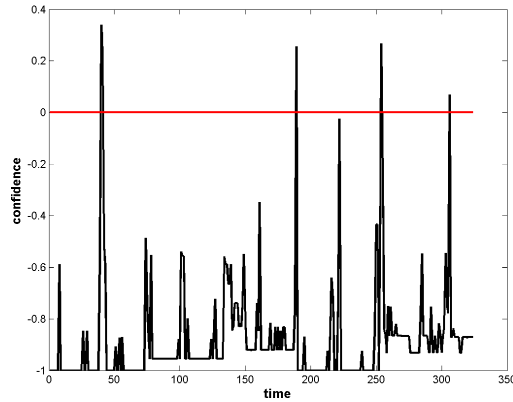
First, we show how a single detector for a single block is evaluated. Therefore, the response (confidence) of the classifier is plotted over time which s shown in Figure 5. It can be seen that the response is increased whenever a person or a part of a person is present in the patch. A detection is reported if the confidence is above some threshold which is usually set to zero. As can be seen from Figure 5 only persons are detected . In contrast, when simply applying background subtraction (temporal median filter) less accurate results are obtained as shown in Figure 6!

Next, we compare the proposed grid-based detector approach to other approaches by analyzing the precision-recall curves. For the grid-based detectors the evaluation was performed on-line; all detection that are reported during the evaluation/updates procedure are included directly into the statistics. For other methods a pre-trained classifier was evaluated on the test sequence. From Figure 7(a) it can be seen that the Dalal and Triggs performs worst among the tested methods. This is not amazing since a generic detector does not include any scene information. In contrast, the Conservative Learning detector yields high recalls producing only a small number of false positives. But a similar performance can be obtained by applying our simple grid-based approach.

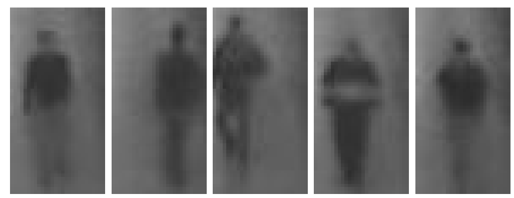
Finally, we want to demonstrate that the trained grid-based classifier is fast adapting to the current environment and never drifts away. Therefore, Figure 7(b) shows the precision-recall curves over time. It can be seen that 50 frames (i.e., updates) are enough to reach a very good performance which is unchanged during the whole experiment.

4.3. Caviar and PETS 2006

To show that the proposed method yields competitive results compared to state-of-the-art methods we applied the grid-



(a)



(b) 41 (c) 189 (d) 222 (e) 255 (f) 306

Figure 5: Confidence of a specific grid-based detector over time (a) and the corresponding images (b)-(f).

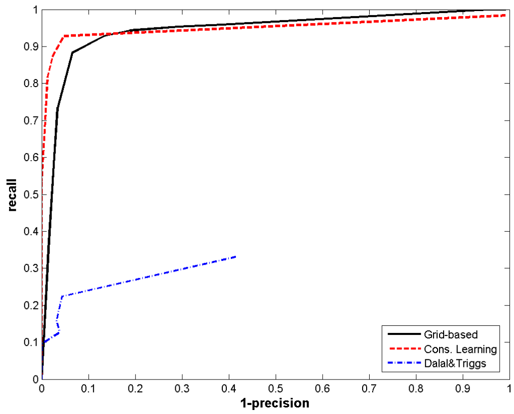
Figure 6: Results for the proposed method (first row) in contrast to a plain background subtraction method (second row).

based person detector on two publicly available datasets, i.e., the *Caviar* dataset[4] and the *PETS 2006* dataset[5].

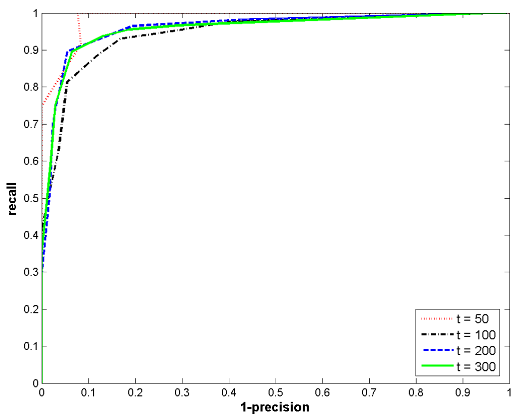
The *Caviar* data set consists of sequences of size 384×288 showing a corridor in a shopping mall. For our ex-

[4] `http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/`, August 29, 2007.

[5] `http://www.pets2006.net`, August 29, 2007.



(a) Comparisons of different detectors.

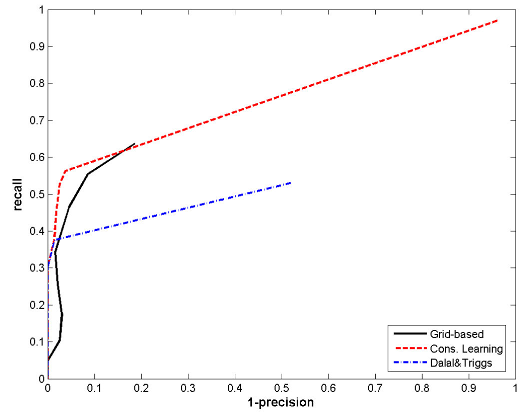


(b) RPC over time for the grid-based approach.

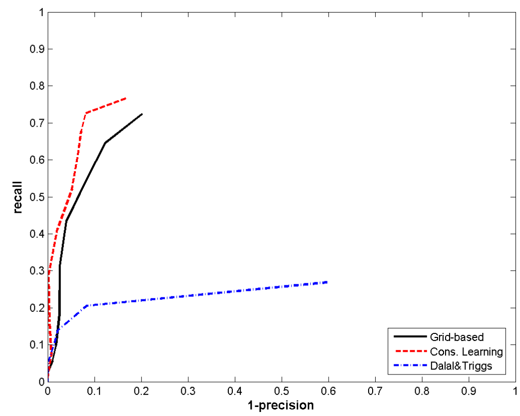
Figure 7: RPC for the Toy Example test sequence. (a) The proposed simple grid-based method reaches results that are obtained by state-of-the-art methods that use sophisticated update strategies. (b) The the grid-based approach benefits from the simplified problem which allows to use a fixed update strategy. Thus, it adapts fast and stable over time with hight performance.

periments we selected one sequence that contains a lot of persons (*ShopAssistant2cor*) and reduced the frame-rate by a factor 10 (370 frames in total). Since the original ground-truth also includes persons that are represented by body-parts only, e.g., by a hand or by a foot, the ground-truth was slightly modified such that only persons are included that are fully visible.

The *PETS 2006* database consists of sequences that show the concourse of a train station from four different views. For our experiments we selected a sequence showing a frontal view (*Dataset S5 (Take 1-G), Camera 4*) and used only every 10th frame (308 frames in total). Due to the high resolution (720×576) and the camera angle the size of the persons varies from approximative 50×100 to 200×100. For evaluation purposes a ground-truth was manually annotated for this sequence.



(a) Caviar.



(b) PETS 2006.

Figure 8: RPC for the public available sequences.

Like in the previous section for both sequences we have generated the precision-recall curves. The results are shown in Figure 8. Since the sequences are harder, i.e., they contain sitting persons, partly occluded persons and persons that occlude each other, compared to the toy example the recall is decreased for all methods. But form the precision-recall curves we see the same trend as for the toy example.

Finally, some representative frames obtained by the proposed detector on all three datasets are depicted in Figure 9. It can be seen that persons are detected independently of their appearance and scale which are varying, especially for the PETS 2006 data set.

5. Summary and Conclusions

We presented a simple grid-based person detector. The main idea is to sub-divide the input images into small overlapping blocks and to train and to maintain a person detector on-line for all of these patches. Since the task of each detector is to detect a person in only one specific patch and at a specific time the complexity of the person detection task is

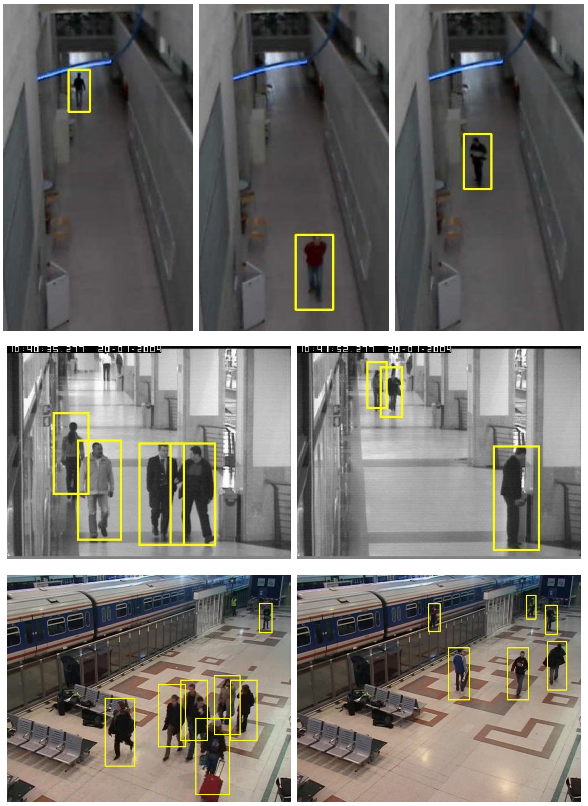


Figure 9: Detection results of our proposed grid-based pedestrian detector on the sequences Toy Example (first row), Caviar (second row) and PETS 2006 (third row).

significantly reduced. Hence, we can apply quite simple and fixed update rules for updating the classifiers. This keeps the classifier stable and limits the drifting problem. In the worst case scenario (a person stands for an extended period of time at the same location) only a certain time interval is affected but the classifier is able to recover completely. This is an essential property for practical applications which run for a long time (24 hours a day, 7 days a week).

In fact, we maintain a small set (as small as one) of positives samples whereas the negative samples are directly drawn from the image sequence. In the experiments we compared the proposed method to state-of-the-art methods. Since we obtain competitive detection results we showed that for a specific surveillance task even a less sophisticated approach would yield comparable results.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *PAMI*, 26(11):1475–1490, 2004.

[2] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. COLT*, pages 92–100, 1998.

[3] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Proc. CVPR*, volume 1, pages 10–17, 2005.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, volume 1, pages 886–893, 2005.

[5] D. M. Gavrila. Multi-feature hierarchical template matching using distance transforms. In *Proc. ICPR*, volume 1, pages 439–444, 1998.

[6] H. Grabner and H. Bischof. On-line boosting and vision. In *Proc. CVPR*, volume 1, pages 260–267, 2006.

[7] H. Grabner, P. Roth, M. Grabner, and H. Bischof. Autonomous learning a robust background model for change detection. In *Proc. PETS*, pages 39–46, 2006.

[8] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Proc. WS on Statistical Learning in Computer Vision*, 2004.

[9] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using co-training. In *Proc. ICCV*, volume 2, pages 626–633, 2003.

[10] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. ECCV*, volume 1, pages 128–142, 2004.

[11] V. Nair and J. Clark. An unsupervised, online learning framework for moving object detection. In *Proc. CVPR*, volume 2, pages 317–324, 2004.

[12] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proc. ICCV*, pages 555–562, 1998.

[13] R. Pflugfelder and H. Bischof. Online auto-calibration in man-made worlds. In *Digital Image Computing: Technqiues and Applications*, pages 519 – 526, 2005.

[14] P. Roth, H. Grabner, D. Skočaj, H. Bischof, and A. Leonardis. On-line conservative learning for person detection. In *Proc. VS-PETS*, 2005.

[15] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. In *Proc. CVPR*, pages 1–8, 2007.

[16] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. CVPR*, volume 2, pages 246–252, 1999.

[17] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *Proc. CVPR*, pages 1–8, 2007.

[18] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, volume I, pages 511–518, 2001.

[19] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. ICCV*, volume 2, pages 734–741, 2003.

[20] B. Wu and R. Nevatia. Improving part based object detection by unsupervised, online boosting. In *Proc. CVPR*, pages 1–8, 2007.

[21] B. Wu and R. Nevatia. Simultaneous object detection and segmentation by boosting local shape feature based classifier. In *Proc. CVPR*, pages 1–8, 2007.