On-line Inverse Multiple Instance Boosting for Classifier Grids

Sabine Sternig, Peter M. Roth, and Horst Bischof

Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{*sternig,pmroth,bischof*}*@icg.tugraz.at*

Abstract

Classifier grids have shown to be a considerable choice for object detection from static cameras. By applying a single classifier per image location the classifier's complexity can be reduced and more specific and thus more accurate classifiers can be estimated. In addition, by using an on-line learner a highly adaptive but stable detection system can be obtained. Even though long-term stability has been demonstrated such systems still suffer from short-term drifting if an object is not moving over a long period of time. The goal of this work is to overcome this problem and thus to increase the recall while preserving the accuracy. In particular, we adapt ideas from Multiple Instance Learning (MIL) for on-line boosting. In contrast to standard MIL approaches, which assume an ambiguity on the positive samples, we apply this concept to the negative samples: Inverse Multiple Instance Learning. By introducing temporal bags consisting of background images operating on different time scales, we can ensure that each bag contains at least one sample having a negative label, providing the theoretical requirements. The experimental results demonstrate superior classification results in presence of non-moving objects.

Keywords:
Classifier Grids, On-line Learning, AdaBoost, Object Detection, Multiple Instance Learning

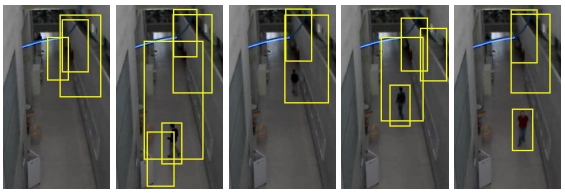
1. Introduction

The first step in many computer vision applications is to identify the objects-of-interest (object detection). The most prominent approach is to apply a sliding window technique (e.g., Dalal and Triggs, 2005; Nair and Clark, 2004; Felzenszwalb et al., 2008; Viola et al., 2003). Each patch of a given image is tested whether it is consistent with a previously estimated model or not, and finally all consistent patches are reported. Typically, the goal of such methods is to build a generic model that is applicable for all possible scenarios and tasks (e.g., Leibe et al., 2008; Felzenszwalb et al., 2008; Dalal and Triggs, 2005).

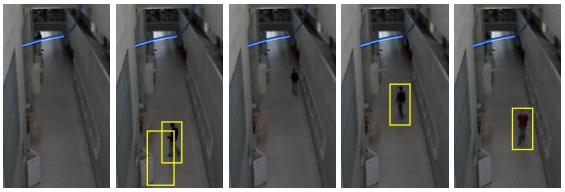
However, as can be seen from Figure 1(a), even if trained from a very large number of training samples such generic detectors ("broad application") often fail in practice. Since not all variability, especially for the negative class (i.e., all possible backgrounds), can be captured this results in a low recall and an insufficient precision. Assuming a stationary camera, which is a reasonable constraint for most applications, using scene specific information can help to reduce the number of false alarms (e.g., Hoiem et al., 2006). As can be seen from Figure 1(b) this can dramatically improve the overall performance of a generic detector. To further improve the classification results specific classifiers ("narrow applications") can be applied, which are designed to solve a specific task (e.g., object detection for a specific setup). In fact, to train such classifiers less training data is required and for the particular task they are usually better in terms of accuracy and efficiency (Levin et al., 2003; Wu and Nevatia, 2007a; Roth et al., 2005). This is illustrated in Figure 1(c).

To further improve the classification power and to further reduce the number of required training samples an adaptive classifier using an on-line learning algorithm can be applied (Nair and Clark, 2004; Javed et al., 2005; Wu and Nevatia, 2007a). Thus, the system can adapt to changing environments (e.g., changing illumination conditions) and these variations need not to be handled by the initial model. In fact, in this way the complexity of the problem is reduced and a more efficient classifier can be trained.

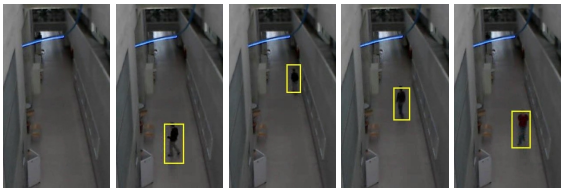
Adaptive systems, however, have one main disadvantage: new unlabeled data has to robustly be included



(a) Fixed Detector.



(b) Scene-specific Detector.



(c) Adaptive Detector.

Figure 1: Changing environmental conditions like lightning changes or changes of objects in the background can not be handled by a fixed model. This requires an adaptive (scene specific) system.

into an already built model. Typical approaches are self-training (e.g., Rosenberg et al., 2005; Li et al., 2007), co-training (e.g., Blum and Mitchell, 1998; Levin et al., 2003), semi-supervised learning (e.g., Goldberg et al., 2008), or the application of oracles[1] (e.g., Nair and Clark, 2004; Wu and Nevatia, 2005). Semi-supervised methods, however, are often biased by the prior and thus only a "limited" information gain can be achieved whereas oracles often provide too less new information. Self- or co-training suffer from the problem that the theoretical constraints can not be assured in practice or that they rely on a direct feedback of the current classifier - both resulting in unreliable classifiers.

More specific and thus more efficient classifiers avoiding these problems can be trained using classifier grids (e.g., Grabner et al., 2007; Stalder et al., 2009; Roth et al., 2009). In contrast to a sliding window technique, where one classifier is evaluated on different image positions, the main idea of classifier grids is to train a separate classifier for each image location. Thus, the complexity of the classification task that has to be handled by a single classifier is dramatically reduced. Each classifier has only to discriminate the object-of-interest from the background at one specific location in the image. By using on-line classifiers the system is able to adapt to changing environmental conditions, which further reduces the required complexity of the classifiers.

Adaptive approaches, in general, suffer from the drifting problem, i.e., due to wrong updates the system starts to learn something completely different degrading the classification performance. To avoid drifting in classifier grids (Roth et al., 2009) applied fixed update strategies. In particular, the negative updates for a grid classifier are generated from the corresponding image patch, whereas the positive representation was pre-trained and kept fix. These update strategies ensure "long-term" stability, i.e., the classifier cannot get totally degenerated. In fact, a classifiers that was trained using wrongly labeled samples would recover within a certain time interval, which we will refer to as "short-term" drifting. This might be the case if an object stays at the same position over a longer period of time and the foreground information is used to model the negative class.

In this work, we address the problem of short-term drifting by incorporating temporal information and replacing the fixed update strategy by a multiple instance learning (MIL)-based approach. In particular, we introduce temporal bags, containing patches of background models operating on different timescales, for each grid element assuming that each bag consists of at least one correctly labeled sample. Since in our case the positive samples are well defined and the ambiguity results from the negative samples, we have to adapt the original MIL concept for our purpose. The experimental results clearly demonstrate the benefits of the proposed method. Especially compared to existing approaches non-moving objects can be handled considerable better, increasing both, the recall and the precision.

The rest of the paper is organized as follows. First, in Section 2, we review the idea of classifier grids. Next, we introduce inverse multiple instance learning for classifier grids in Section 3. In Section 4, we give an experimental evaluation of the proposed approach. Finally, we summarize and conclude the paper in Section 5.

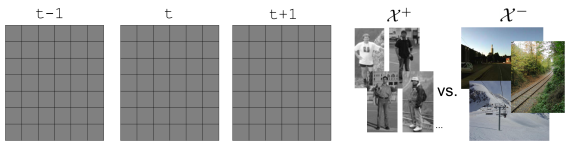
2. Classifier Grids

In the following, we give an short overview of object detection from static cameras, review the main ideas classifier grids, and discuss their practical applicability under real-world conditions.

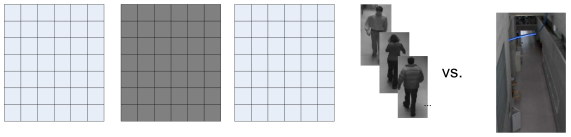
[1] An oracle can be considered a classifier, even at a low recall rate, having a high precision, which can be used to generate new training samples.

2.1. Object Detection from Static Cameras

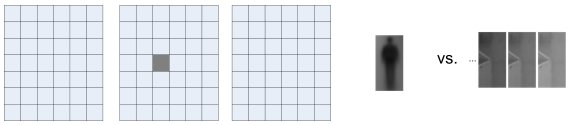
Even though for most object detation scenarios a stationary camera can be assumed, this constraint, which could help to drastically improve the classification performance, has been only of limited interest (e.g., Hoiem et al., 2006; Roth and Bischof, 2008; Wu and Nevatia, 2007b; Nair and Clark, 2004). However, there are three different concepts for object detection from stationary cameras: (a) fixed models which are trained off-line, (b) adaptive models which are trained on-line, and (c) classifier grids. These are illustrated in Figure 2, where the dark highlighted regions 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.



(a) Fixed general detector.



(b) Scene-specific detector.



(c) Grid-based detector.

Figure 2: Overview of different concepts for object detection from static cameras and the corresponding training sets: (a) fixed detector, (b) scene specific detector, and (c) grid-based detector. The gray blocks highlight the regions in both, time and location, where the classifier has to perform well.

In general, a 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 is used to train a detector. In the first case (fixed detector), which is illustrated in Figure 2(a), the training set $\mathcal{X}$ is fixed and a classifier is trained using an off-line training algorithm. 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.

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 from the background of a specific scene. Thus, the variability of the background as well as the number of required training samples is reduced as illustrated in Figure 2(b).

2.2. Classifier Grid and Fixed Update Rules

The main idea of classifier grids (Grabner et al., 2007; Roth et al., 2009) is to exploit the prior knowledge, that the camera is fixed. By using this information, the whole detection task can be simplified by sampling the input image into by using a fixed highly overlapping grid (both in location and scale), where each grid element $i = 1, \ldots, N$ corresponds to one classifier C_i. This is illustrated in Figure 3. Thus, the classification task that has to be handled by one classifier C_i is reduced to discriminate the background of the specific grid element from the object-of-interest. Moreover, stationary cameras allow to pre-estimate the ground-plane, which further helps to reduce the number of classifiers within the classifier grid. Due to this simplification less complex classifiers can be applied. In particular, the grid-based representation is well suited for compact on-line classifier, which can be evaluated and updated very efficiently.

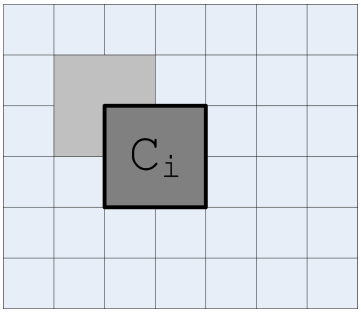


Figure 3: Concept of grid-based classification: a highly overlapping grid is placed over the image, where each grid element corresponds to a single classifier.

The main problem of adaptive system is to robustly incorporate unlabeled data from the scene, which may lead to the "drifting problem". By talking about "drifting" the problem is that wrong updates may completely destroy the classifier. To overcome this problem, at time t fixed updates (Grabner et al., 2007) can be applied for updating a classifier $C_{i,t-1}$. Given a set of representative 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 by definition. The probability that an object is present in a patch $\mathbf{x}_i$ is given by

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

where $\#p_i$ is the number of objects 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 $P(\mathbf{x}_i = \text{object})$). The probability of a wrong update for this particular image patch is indeed very low.

2.3. Discussion

The fixed update strategies cause three main problems. First, even if the positive information is kept fixed, positive updates are required. Second, no new positive information can be acquired. Third, still wrong negative updates might occur leading to short-term drifting.

The first problem was addressed in (Roth et al., 2009), where the main idea was to further increase the stability and to speed up the computation by a combination of two generative models in parallel: a pre-trained model for the positive class and an adaptive model for the negative class. The pre-trained model for the positive class can be calculated in an off-line manner. By using off-line boosting for feature selection this gives the additional advantage, that the classifier is initialized by features well suited for the task of interest in contrast to a random initialization of a on-line classifier. Since well suited features are selected within the classifier, one further has the advantage that the classifier size can be reduced compared to randomly initialized classifiers, where a larger classifier size is required for a good classification result. The strong positive prior inhibits fast temporal drifting while the negative updates during runtime ensure the required adaptivity. Moreover, since the positive model is kept fix, the number of required updates is reduced.

The second problem was addressed in (Stalder et al., 2009) and in (Sternig et al., 2010b). (Stalder et al., 2009) introduced context-based classifier grids to extract additional positive information from a specific scene. This context information is gained through three different ways: a fixed detector, a tracker and 3D-context information. The authors showed that the recall can be drastically increased, but on the expense of the precision. In contrast, in (Sternig et al., 2010b) we proposed to use a co-training approach (*Classifier Co-Grids*) in combination with a robust on-line learner. The robust on-line learner keeps two seperate models for the positive class and two separate models for the negative class. For both, the positive and the negative class one model is off-line pre-trained and kept fixed during runtime, while one model is adapted. This combination of an off-line pre-trained model with an on-line adapted model within a robust on-line learner allows to incorporate scene specific positive information (i.e., the recall can be increased), while still preserving the accuracy of our system.

However, if too many wrong updates are performed, a foreground object may grow into the background class and the detector fails (i.e., it generates a miss). Even though the detector recovers quickly – within a few frames (short time drifting), the goal would be to avoid this problem. Thus, this open problem is addressed in the following by using the idea of Inverse Multiple Instance Learning.

3. Inverse MIL for Classifier Grids

Even though the updates generated by the fixed rules are correct most of the time, they might be wrong causing the classifier to drift within a certain time interval. Especially, if an object is not moving over a long period of time, foreground information is used to perform negative updates, which causes the positive information to be temporally unlearned. Since this can be seen in the context of ambiguous labeled samples, Multiple Instance Learning could help to deal with this problem. Thus, in the following we first review the main ideas of MIL, derivate an inverse on-line MILBoost algorithm, and show how it can be applied in the context of classifier grids. In particular, we build on the boosting approach presented in (Roth et al., 2009), which already ensures the long-term robustness.

3.1. Multiple Instance Learning

Multiple-instance learning (MIL) was first introduced by (Dietterich et al., 1997). It is a machine learning paradigm for dealing with ambiguously labeled data. Thus, there has been a considerable interest and various different approaches have been proposed. Most of these approaches are based on popular supervised learning algorithms such as SVM (Andrews et al., 2003) or boosting (Viola et al., 2005), that are adapted in order to incorporate the MIL constraints.

In contrast to supervised learning algorithms, where each sample (instance) is provided a label, in multiple-instance learning the training samples are grouped into bags $B_i \subset \mathbb{R}^d, i = 1, \ldots, N$. Each bag consists of an arbitrary number of instances: $B_i = \{x_{1i}, x_{2i}, \ldots, x_{m_i i}\}$. Negative bags B_i^- are required to consist only of negative instances, whereas for positive bags B_i^+ it has only to be guaranteed that they contain at least one positive instance. There are no further restrictions to the non-positive instances within the positive bag B_i^+, they might not even belong to the negative class.

The task now is to learn either a bag classifier $f : B \rightarrow \{-1, 1\}$ or an instance classifier $f : \mathbb{R}^d \rightarrow \{-1, 1\}$. However, a bag classifier can follow automatically from instance prediction, e.g., by using the *max* operator over posterior probabilities over the instances p_{ij} within the i^{th} bag: $p_i = \max_j\{p_{ij}\}$.

3.2. On-line Inverse MILBoost

In general, the goal of boosting is to estimate a strong classifier

$$H(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j h_j(\mathbf{x}) \tag{4}$$

by a linear combination of N weak classifiers $h_j(\mathbf{x})$. In particular, we build on Babenko et al. (Babenko et al., 2009) and use a different loss function, optimizing the binary log likelihood over bags in form of

$$log \quad \mathcal{L} = \sum_i \left(y_i log\, p(y_i) + (1 - y_i) log\, (1 - p(y_i))\right), \tag{5}$$

where the instance probability can be estimated using a sigmoid function

$$p(y|x) = \sigma(H(x)) = \frac{1}{1 + e^{-H(x)}}, \tag{6}$$

which requires a gradient descent in function space. The bag probability $p(y|B)$ is modeled by the Noisy-OR (NOR) operator:

$$p(y_i|B_i) = 1 - \prod_{j=1}(1 - p_(y_i|x_{ij}). \tag{7}$$

However, since in the classifier grid scenario the positive samples are well defined (positive samples are hand labeled) and the ambiguity concerns only the negative samples (comming directly from the scene without labeling), the original MIL idea has to be adapted. Thus, the negative bags B_i^- would need to contain only one negative example whereas the positive bag B_i^+ consists only of positive examples:

$$\forall x_{ij}^+ \in B_i^+ : y(x_{ij}^+) = 1 \tag{8}$$

$$\exists x_{ij}^- \in B_i^- : y(x_{ij}^-) = -1\ . \tag{9}$$

In order to correctly calculate the loss $\mathcal{L}$ by inverting the problem, we have to switch the labels between the positive and the negative class (*inverse MIL*). This causes to focus on examples that are more likely to be correct negative examples, which directly fits to our problem.

3.3. IMIL in Classifier Grids

Building on (Roth et al., 2009) the model describing the object class (the positive model) is pre-calculated off-line and only negative updates are performed. Thus, we can neglect the positive bags. To generate the negative bags, we can collect a stack of input images from the image sequence over time, which we refer to as "temporal bag". Having a large stack assures that the assumption for the negative bag containing at least one negative sample is mostly valid, since the probability that an object stays at one specific location over a longer period of time is very low (Sternig et al., 2010a).

Collecting a large stack of input images is adversarial for the runtime behavior. In order to avoid a large stack of input images within the temporal bag we use a small set of background images which operate on different time scales, which means that they are updated in different time intervals. Any kind of background model can be used. However, in our case we apply the approximated median background model (McFarlane and Schofield, 1995). Since these background models are updated in different time intervals even a small number ensures that the temporal bags fulfill the MIL constraints. To give enough adaptivity to chaning illumination conditions we have background models which are updated in small time intervals, while to avoid objects staying at the same position for a while to become part of the background we have background models updated in long time intervals. Hence, the multiple instance learning property of inherently dealing with ambiguity in data can be exploited for improving the classifier grid approach and avoiding short-term drifting.

4. Experimental Results

To demonstrate the benefits of the proposed approach, we run five experiments considering two tasks,

namely pedestrian and car detection. We first give an illustrative comparison between the original grid approach (e.g., Roth et al., 2009) and the proposed method. Then, using a publicly available pedestrian dataset we show that by the IMIL grid approach state-of-the-art (or even better) detection results can be obtained. Finally, we selected a number of datasets (pedestrians and cars) containing objects which are not moving over a long period of time. This causes short-term drifting in existing classifier grid approaches (e.g., Roth et al., 2009), which is in particular the problem addressed within this paper. From all experiments the benefits of the proposed methods are clearly visible.

For all experiments on pedestrian detection we use classifiers consisting of 30 selectors, where each selector consists of a set of 30 weak classifiers. For the car detection experiment we use classifiers consisting of 50 selectors, each of them containing 30 weak learners. As weak classifiers we use simple decision stumps over the feature responses of Haar-like features. To increase the robustness of the negative updates, we collect a stack of four background images, operating on four different timescales, which are updated every second frame, every 50-th frame, every 100-th frame, and every 150-th frame.

For practical applications, it is not necessary to update the system with every input frame (typically there is a trade-off between runtime and adaptivity to changing environments). However, to demonstrate the benefits and the robustness of our approach, i.e., the avoidance of temporal drifting, we update each classifier within the classifier grid with every single input frame. The overlap of the grid elements within the classifier grid is set to 70% for the pedestrian sequences and to 85% for the car sequence. For calculating the Recall-Precision-Curves (RPC) a detection is counted as true positive if it fulfills the overlap criterion (Agarwal et al., 2004), where a minimal overlap of 50% is required.

4.1. IMIL Behavior Analysis

First of all, we want to illustrate the benefits of the proposed approach compared to the CG approach by considering the particular case that an object (i.e., a person) is not moving for a longer time. We picked out a sub-sequence of the longterm experiment (see Section 4.5) where one person is standing at the same position over 450 frames and analyze the confidences at one specific position in the image. The confidence for both, the proposed approach and the CG approach as well as the ground truth are shown in Figure 4. One can clearly see that the confidence for the proposed approach stays the same due to the correct updates with our inverse multiple instance learning strategy while the wrong updates with the current input image of the CG approach lead to decreasing confidence over time. Moreover, it can also be seen that for negative class, i.e., the background the confidence is much lower.

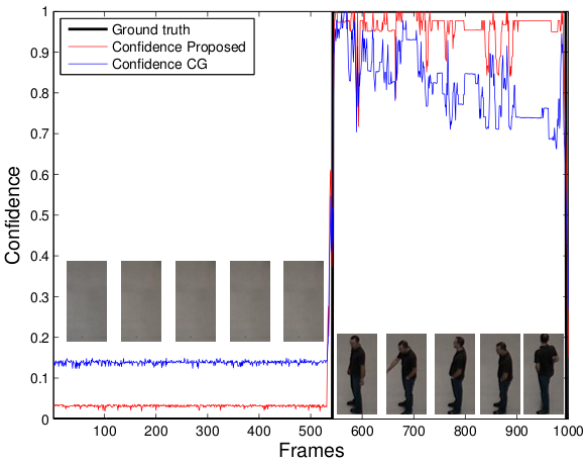


Figure 4: Confidence values for the proposed approach and the CG approach for a typical scenario: left - background, right - person standing on the same position for a longer period of time.

4.2. PETS 2006

For this experiment we used a sequence from the publicly available *PETS 2006* dataset consisting of 308 frames (720x576 pixels), which contains 1714 pedestrians. We compare our approach to other state-of-the-art person detectors, namely the deformable part model (Felzenszwalb et al., 2008) (FS) and the Histograms of Oriented Gradients approach (Dalal and Triggs, 2005) (DT). Both approaches use fixed off-line trained classifier and are based on the sliding window technique. In addition, we compared our method to the classifier grid (CG) approach (Roth et al., 2009). Since the classifier grid approaches use ground plane information to generate the grids we removed all false positives for the sliding window based detectors which are smaller than 75% or larger than 125% of the groundtruth size in order to guarantee a fair comparison.

The results are shown in Figure 5, where it can be seen that the proposed approach clearly outperforms the generic detectors as well as the original Classifier Grid approach, which can be considered a baseline for the proposed method. In addition, in Table 1 we give the recall and precision for the best F-Measure value and show some qualitative results in Figure 6. In this case having an image size of 720x567 a classifier overlap of 70% results in 785 classifiers. Using a non-optimized,

not parallelized implementation run on a standard PC, thus a computation time of approximative 1.5 seconds per frame is required.

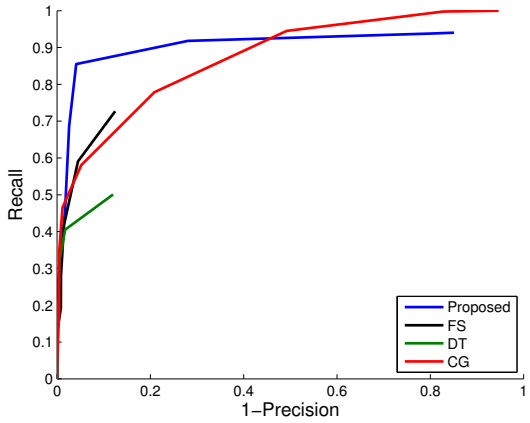


Figure 5: Recall-Precision Curves for *PETS 2006* sequence for different state-of-the-art detectors compared to the proposed approach.

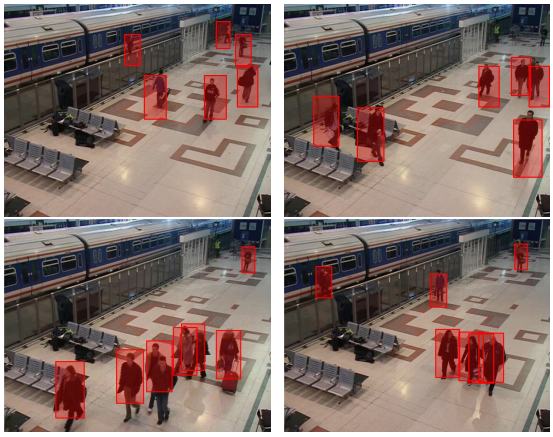


Figure 6: Illustrative detection results of our approach on the *PETS 2006* sequence.

	Recall	Precision	F-Measure
Proposed	0.86	0.96	0.90
FS	0.73	0.88	0.79
DT	0.50	0.88	0.64
CG	0.78	0.79	0.78

Table 1: Comparison of Recall and Precision for the best F-Measure value for different approaches for *PETS 2006*.

4.3. Corridor Sequence

To demonstrate the benefits of our approach in presence of non-moving objects compared to existing classifier grid detectors, we generated a test sequence showing exactly this problem: *Corridor* sequence. The sequence showing a corridor in a public building consists of 900 frames (640x480) containing 2491 persons, which are staying at the same position over a long period of time. The results obtained by the proposed approach and the CG method are shown in Figure 7.

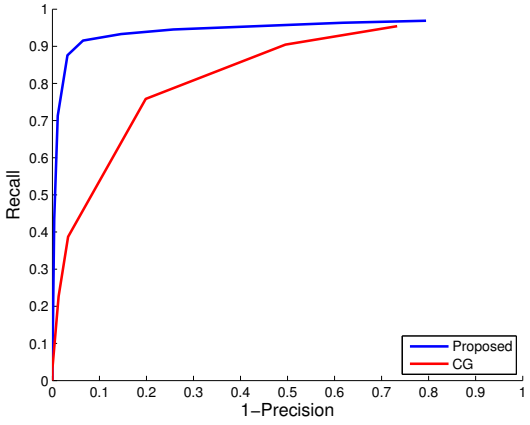


Figure 7: Recall-Precision Curves for the *Corridor* sequence for the original Classifier Grid and the proposed approach.

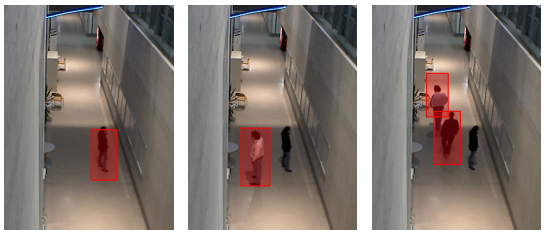
Since due to the IMIL formulation we get rid of short-term-drifting, the recall (at a reasonable precision level) can be significantly improved. This is also illustrated in Figure 8, where the first row shows detection results of the original classifier grid approach, whereas the second row shows detection results using the proposed inverse multiple-instance learning. It can clearly be seen that the person on the right side, standing at the same position over 175 frames, is detected by the proposed approach whereas it is not in the other case. In addition, Table 2 shows the recall and precision for the best F-Measure value.

	Recall	Precision	F-Measure
Proposed	0.92	0.93	0.92
CG	0.76	0.80	0.78

Table 2: Comparison of Recall and Precision for the best F-Measure value for the original CG and the proposed approach on the *Corridor Sequence*.

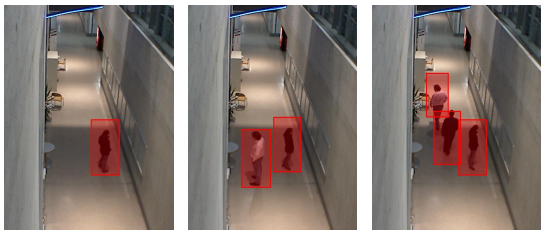
4.4. Vehicle Sequence

To demonstrate that our approach is not limited to pedestrian detection, we additionally evaluate it on a sequence showing vehicles on a highway: *Vehicle* sequence. This sequence consists of 500 frames



Frame 99 Frame 164 Frame 274

(a) Classifier Grid (Roth et al., 2009).



Frame 99 Frame 164 Frame 274

(b) Proposed approach.

Figure 8: Temporal information incorporation by MIL avoids short-term drifting. The original classifier grid approach (first row) temporary drifts after about 60 frames whereas the proposed approach (second row) avoids temporal drifting even after more than 170 frames.

(720x576), containing 2375 cars. One car broke down within this sequence and is standing at the same position for 400 frames.

The Recall-Precision curves, again for the proposed and the original CG approach, are shown in Figure 9. Again it can be seen that compared to the baseline approach the detection performance can be noticeable improved. Additionally, illustrative detection results for this scenario are shown in Figure 10. Table 3 shows the recall and precision for the best F-Measure value.

	Recall	Precision	F-Measure
Proposed	0.99	0.92	0.95
CG	1.00	0.85	0.92

Table 3: Comparison of recall and precision for the best F-Measure value for the original CG and the proposed approach on *Vehicle* sequence.

4.5. Longterm Pedestrian Detection

Finally, we want to demonstrate that the longterm stability, which was already shown for the original grid approach in (Roth et al., 2009), also holds for the IMIL extension. Since the proposed idea builds on the same fixed update strategy the overall longterm stability is still ensured. Thus, we have to show that also short-term drifting can be avoided when running for a longer period of time.

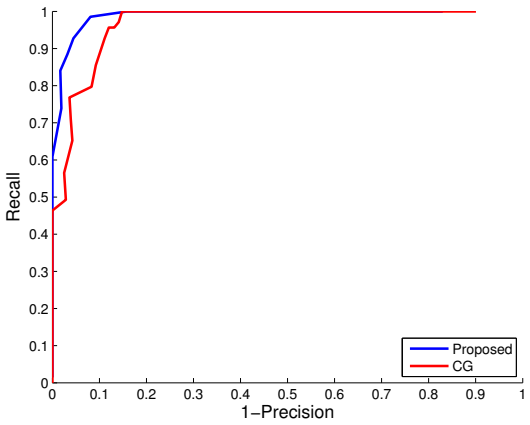


Figure 9: Recall-Precision Curves for the *Vehicle* sequence, containing objects that are not moving over a long period of time.

Figure 10: Illustrative detection results on the *Vehicle* sequence.

For that purpose, we recorded a sequence consisting of 435.000 frames captured over 24 hours with a frame rate of approximately five frames per second, showing the same difficulties as described in Section 4.3 (i.e., containing people standing on the same position over a longer period of time). To demonstrate the robustness over time we hand-labeled three sub-sequences at different points in time: one right at the beginning, starting at frame 3500, one in the middle of the sequence, starting at frame 312.000, and one close to the end, starting at frame 416.000. All sub-sequences have a length of 2000 frames and contain 956, 940 and 603 pedestrians, respectively.

The thus obtained Recall-Precision-Curves shown in Figure 11 demonstrate that the performance stays the same even after more than 415.000 updates. The minor differences can be explained by the different absolute

number of true positives, which directly influences the impact of false positives to the precision.

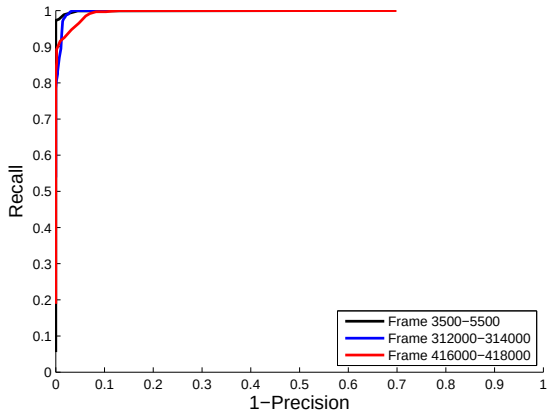


Figure 11: Recall-Precision Curves for longterm sequence consisting of 435.000 frames evaluating the performance of the proposed approach on three different points in time.

4.6. Discussion

From the results presented in this section it can be seen that classifier grids, in general, provide a considerable alternative to typical sliding window approaches when run on static cameras. In particular, only a small number of positive samples (approximative 500) is required to get a meaningful model for the positive class. This has to be done only once and can be re-used for different scenarios. Using more samples would not degrade the classification results, but there is also no benefit since the strength of the method mainly results from the negative data captured during runtime. Since the model is adapted online, minor movements or even a large displacement of the the camera could be compensated as long as the geometry of the ground plane is not changed too much.

As we already showed in (Roth et al., 2009) that the approach is robust even when running in a real-world 24/7 setup, the goal of this paper was to address the problem of short-time drifting if objects are not moving for a longer period of time. This effect was illustrated in Section 4.1. also showing that the proposed inverse multiple instance approach could cope with this problem. This also can be seen from the other experiments (Section 4.2-4.4) which were run for person an car detection: the accuracy of the detection results can be increased. Moreover, in Section 4.5 we showed that the robustness of the original approach is preserved and that the method yield excellent detection results even if the system is updated thousands of times. As drawback, however, the MIL implementation increases the runtime. But this could be compensated exploiting the highly parallel structure of the approach.

5. Conclusion

Having a stationary camera, which is a reasonable assumption for many detection tasks, classifier grids can be applied instead of sliding window approaches. However, due to fixed update strategies, using the current input image to update the negative representation, non-moving objects cause the system to drift temporary, even though it is able to recover later on. To cope with this specific problem, we proposed to adopt multiple-instance learning (MIL), a well known machine learning approach for handling ambiguously labeled positive samples. However, since in our case the ambiguity concerns the negative samples, we modified the original multiple-instance learning idea (inverse MIL). We adapted on-line MILBoost (Babenko et al., 2009) to fit to our problem. In particular, as in (Roth et al., 2009) we kept the positive representation fixed an generated a bag of negative samples from an estimated background model. The experimental results, demonstrated on three different setups, clearly show that state-of-the-art results can be obtained and that the problem of short-term drifting can be avoided clearly improving the detection performance.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF) under the project MASA (P22299) and the Austrian Research Promotion Agency (FFG) under the project SECRET (821690) and the Austrian Security Research Programme KIRAS.

References

Agarwal, S., Awan, A., Roth, D., 2004. Learning to detect objects in images via a sparse, part-based representation. IEEE Trans. on Pattern Analysis and Machine Intelligence 26 (11), 1475–1490.

Andrews, S., Tsochantaridis, I., Hofmann, T., 2003. Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems. pp. 561–568.

Babenko, B., Yang, M.-H., Belongie, S., 2009. Visual tracking with online mulitple instance learning. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition.

Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In: Proc. Conf. on Computational Learning Theory. pp. 92–100.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Vol. I. pp. 886–893.

Dietterich, T. G., Lathrop, R. H., Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89 (1–2), 31–71.

Felzenszwalb, P., McAllester, D., Ramanan, D., 2008. A discriminatively trained, multiscale, deformable part model. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition.

Goldberg, A. B., Li, M., Zhu, X., 2008. Online manifold regularization: A new learning setting and empirical study. In: Proc. European Conf. on Machine Learning and Knowledge Discovery in Databases. Vol. I. pp. 393–407.

Grabner, H., Roth, P. M., Bischof, H., 2007. Is pedestrian detection really a hard task? In: Proc. IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance. (in conj. ICCV).

Hoiem, D., Efros, A. A., Hebert, M., 2006. Putting objects in perspective. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Vol. II. pp. 2137–2144.

Javed, O., Ali, S., Shah, M., 2005. Online detection and classification of moving objects using progressively improving detectors. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Vol. I. pp. 696–701.

Leibe, B., Leonardis, A., Schiele, B., 2008. Robust object detection with interleaved categorization and segmentation. Int'l Journal of Computer Vision 77 (1–3), 259–289.

Levin, A., Viola, P., Freund, Y., 2003. Unsupervised improvement of visual detectors using co-training. In: Proc. ICCV. Vol. I. pp. 626–633.

Li, L.-J., Wang, G., Fei-Fei, L., 2007. Optimol: automatic online picture collection via incremental model learning. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. pp. 1–8.

McFarlane, N. J. B., Schofield, C. P., 1995. Segmentation and tracking of piglets. Machine Vision and Applications 8 (3), 187–193.

Nair, V., Clark, J. J., 2004. An unsupervised, online learning framework for moving object detection. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Vol. II. pp. 317–324.

Rosenberg, C., Hebert, M., Schneiderman, H., 2005. Semi-supervised self-training of object detection models. In: IEEE Workshop on Applications of Computer Vision. pp. 29–36.

Roth, P. M., Bischof, H., 2008. Machine Learning Techniques for Multimedia. Springer, Ch. Conservative Learning for Object Detectors, pp. 139–158.

Roth, P. M., Grabner, H., Skočaj, D., Bischof, H., Leonardis, A., 2005. On-line conservative learning for person detection. In: Proc. IEEE Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. pp. 223–230, (in conj. ICCV).

Roth, P. M., Sternig, S., Grabner, H., Bischof, H., 2009. Classifier grids for robust adaptive object detection. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition.

Stalder, S., Grabner, H., Gool, L. v., 2009. Exploring context to learn scene specific object detectors. In: Proc. IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance.

Sternig, S., Roth, P. M., Bischof, H., 2010a. Inverse Multiple Instance Learning for Classifier Grids. In: Proc. Intern. Conf. on Pattern Recognition.

Sternig, S., Roth, P. M., Bischof, H., 2010b. Learning of scene-specific object detectors by classifier co-grids. In: Proc. IEEE Int'l Conf. on Advanced Video and Signal-Based Surveillance.

Viola, P., Jones, M. J., Snow, D., 2003. Detecting pedestrians using patterns of motion and appearance. In: Proc. ICCV. Vol. II. pp. 734–741.

Viola, P., Platt, J. C., Zhang, C., 2005. Multiple instance boosting for object detection. In: Advances in Neural Information Processing Systems. pp. 1417–1426.

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

Wu, B., Nevatia, R., 2007a. Improving part based object detection by unsupervised, online boosting. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. pp. 1–8.

Wu, B., Nevatia, R., 2007b. Simultaneous object detection and segmentation by boosting local shape eature based classifier. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. pp. 1–8.