# Inverse Multiple Instance Learning for Classifier Grids

Sabine Sternig, Peter M. Roth, Horst Bischof
*Institute for Computer Graphics and Vision, Graz University of Technology, Austria*
{*sternig,pmroth,bischof*}*@icg.tugraz.at*

***Abstract*—Recently, classifier grids have shown to be a considerable alternative for object detection from static cameras. However, one drawback of such approaches is drifting if an object is not moving over a long period of time. Thus, the goal of this work is to increase the recall of such classifiers while preserving their accuracy and speed. In particular, this is realized by adapting ideas from Multiple Instance Learning within a boosting framework. Since the set of positive samples is well defined, we apply this concept to the negative samples extracted from the scene: Inverse Multiple Instance Learning. By introducing temporal bags, we can ensure that each bag contains at least one sample having a negative label, providing the required stability. The experimental results demonstrate that using the proposed approach state-of-the-art detection results can by obtained, however, showing superior classification results in presence of non-moving objects.**

## I. Introduction

The most prominent approach for object detection is to use a sliding window technique (e.g., [1], [2], [3]). 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. By having a stationary camera, which is a reasonable assumption for many practical scenarios, classifier grids (e.g., [4], [5]) have shown to be a proper alternative to sliding window approaches. 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, which further reduces the required complexity of the classifiers allowing for real-time object detection.

In order to cope with changing environments (e.g., changing illumination conditions, changing backgrounds, ...) the system needs to be adaptive, which requires to incorporate new unlabeled samples. 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. [5] 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. Thus, a classifiers will recover if it was trained using wrongly labeled samples for some time, 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 based approach. In particular, we introduce temporal bags 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 show that in presence of non-moving objects the recall can be increased while the accuracy can still be ensured. In particular, even though not limited to this application, we demonstrate the approach for the task of person detection.

## II. Classifier Grids and Multiple Instance Learning

In the following, we review the ideas of classifier grids and multiple instance learning, which build the base for the proposed approach.

### A. Classifier Grids

The main concept of classifier grids ([4], [5]) is to sample an input image by using a highly overlapping grid, where each grid element $i = 1, \ldots, N$ corresponds to one classifier $C_i$. This is illustrated in Figure 1. To reduce the number of classifiers within the classifier grid the ground-plane is pre-estimated. Thus, the classification task that has to be handled by one classifier $C_i$ can be drastically reduced, i.e., discriminating the background of the specific grid element from the object-of-interest.

To further reduce the classifiers' complexity and to increase the adaptivity, on-line learning methods can be applied, where the updates are generated by fixed rules. For positively updating a grid classifier $C_i$ a fixed pool of positive samples is used; the negative updates are generated directly from the image patches corresponding to a grid element. In general, for estimating the grid classifiers any on-line learning algorithm can be applied, however, on-line boosting has proven to be a considerable trade-off between speed and accuracy [6].

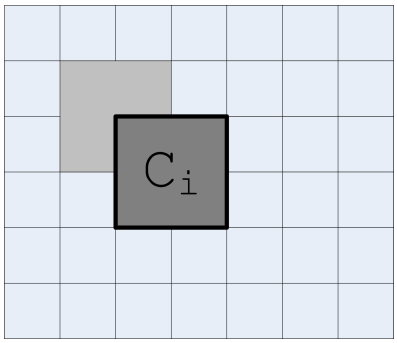



Figure 1. Grid-based classification: a highly overlapping grid of classifiers is placed over the image.

To further increase the stability and to speed up the computation a combination of two generative models can be applied in parallel [5]: a pre-trained model for the positive class and an adaptive model for the negative class, which is still updated using samples from the scene. In this way the strong positive prior inhibits fast temporal drifting while ensuring the required adaptivity.

### *B. Multiple Instance Learning*

Multiple-instance learning (MIL), first introduced by Dietterich et al. [7], is a machine learning paradigm that deals with ambiguously labeled data. In contrast to supervised learning algorithms, where each sample (instance) is provided a label, in multiple-instance learning training samples are grouped in bags $B_i \subset \mathbb{R}^d, i = 1, \ldots, N$, where each bag consist of an arbitrary number of instances: $B_i = \{x_{1i}, x_{2i}, \ldots, x_{m_i i}\}$. Negative bags $B_i^-$ 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 in $B_i^+$, 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, bag classification can follow automatically from instance prediction, e.g., by using the *max* operator $p_i = \max_j\{p_{ij}\}$ over posterior probabilities over the instances $p_{ij}$ within the $i^{th}$ bag. Thus, there have been various multiple-instance learning extensions of popular supervised learning algorithms. In particular, Viola et al. [8] developed a multiple instance boosting algorithm (MILBoost) and applied it to object detection. This algorithm was later on adopted to the on-line domain by Babenko et al. [9] allowing for stable tracking of objects.

## III. 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. Especially, if an object is not moving over a long period of time, also foreground information is labeled as negative and the positive information is temporally unlearned. Since this can be seen in the context of ambiguous labeled samples, Multiple Instance Learning could help to solve this problem.

To avoid short-term drifting, while still preserving the long-term robustness (due to the combination of off-line pre-trained positive distributions $D_j^+$ and on-line estimated negative distributions $D_j^-$), we adapt the boosting approach presented in [5] to the MIL domain. Thus, the goal is to estimate a strong classifier

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

by a linear combination of $N$ weak classifiers $h_j(\mathbf{x})$. Similar to Babenko et al. [9] we use a different loss function, optimizing the binary log likelihood over bags in form of

$$log \quad \mathcal{L} = \sum_i \big(y_i log\, p(y_i) + (1 - y_i) log\, (1 - p(y_i))\big), \tag{2}$$

where the instance probability can be estimated using a sigmoid function

$$p(y|x) = \sigma(H(x)) = \frac{1}{1 + e^{-H(x)}}, \tag{3}$$

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

$$p(y_i|B_i) = 1 - \prod_{j=1} (1 - p_(y_i|x_{ij}). \tag{4}$$

However, since the positive samples are well defined and the ambiguity concerns only the negative samples, the original MIL idea has to be adapted. In our case the negative bags $B_i^-$ would need to contain only one negative example whereas the positive bag $B_i^-$ consists only of positive examples:

$$\forall x_{ij}^+ \in B_i^+ : y(x_{ij}^+) = 1 \tag{5}$$

$$\exists x_{ij}^- \in B_i^- : y(x_{ij}^-) = -1 \tag{6}$$

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. Since only negative updates are performed, we can neglect the positive bags. To generate the negative bags, we 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:

$$P(x = \text{object}) = \frac{\#p_i}{\Delta t} \ . \tag{7}$$

Hence, the multiple instance learning property of inherently dealing with ambiguity in data can be used for improving the classifier grid approach and avoiding short-term drifting.

## IV. Experimental Results

Although not limited to this application, we demonstrate our method on the task of pedestrian detection. In order to demonstrate the benefits of the proposed approach, we show two experiments. In the first experiment we illustrate that we can obtain state-of-the-art person detection results on publicly available datasets. The second experiment shows that by using the proposed approach short-term drifting can be handled considerable better than by existing classifier grid approaches such as [5].

For all experiments we use classifiers consisting of 30 selectors, each of them containing 30 weak learners. As weak classifiers we calculated simple stumps over the feature response of Haar-like features. For calculating the Recall-Precision-Curves (RPC) a detection is counted as true positive if it fulfills the overlap criterion [10], where a minimal overlap of 50% is required.

### A. PETS 2006

For the first 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 of Felzenszwalb et al. [2] (FS) and the Histograms of Oriented Gradients approach of Dalal and Triggs [3] (DT). Both approaches use fixed classifier and are based on the sliding window technique. In addition we compared our method to the classifier grid (CG) approach of Roth et al. [5]. Both classifier grid approaches use ground plane information to generate the grids. Thus, to enable a fair comparison, we removed all false positives for the sliding window based detectors which are smaller than 75% or larger than 125% of the grundtruth size. Since within this sequence there is only one person standing at the same position over a few frames, there is only a slightly improvement compared to the other approaches as shown in the Recall-Precision Curves (RPC) in Figure 2. In addition, qualitative results are shown in Figure 3.

### B. 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.

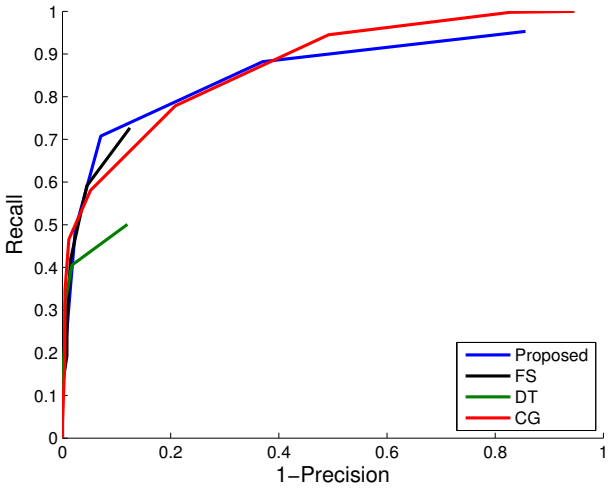



Figure 2. Recall-Precision Curves for PETS 2006 sequence. Our proposed approach reaches slightly better results compared to state-of-the-art object detectors on the publicly available dataset.

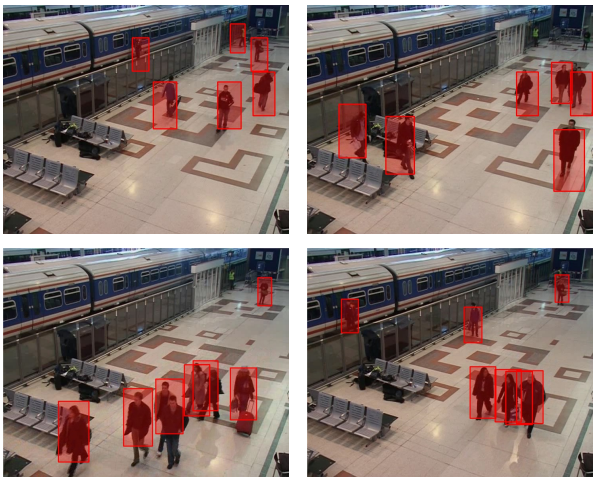


Figure 3. Illustrative results on the PETS 2006 sequence.

The results obtained by the proposed approach and the CG method are shown in Figure 4. Since the temporary drifting can be avoided, it clearly can be seen that the recall is significantly improved. The same behavior can be recognized from Figure 5, which visualizes the difference between the two approaches. The first row show detection results of the original classifier grid approach, whereas in the second row detection results using the proposed inverse multiple-instance learning strategy are illustrated. It clearly can be seen that the person on the right side, standing at the same position over 175 frames, is detected by the proposed approach where it is not detected by the other.

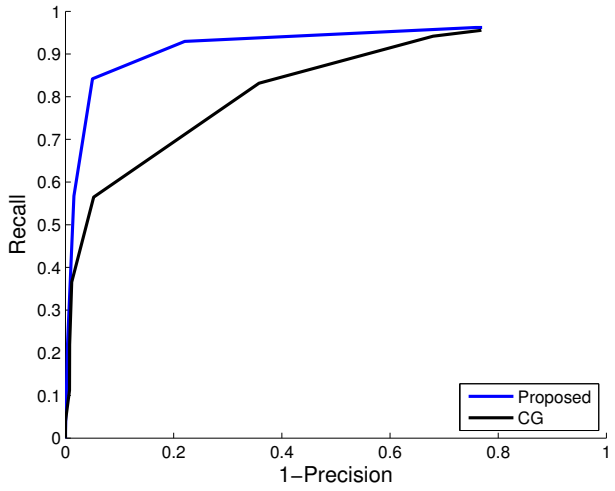
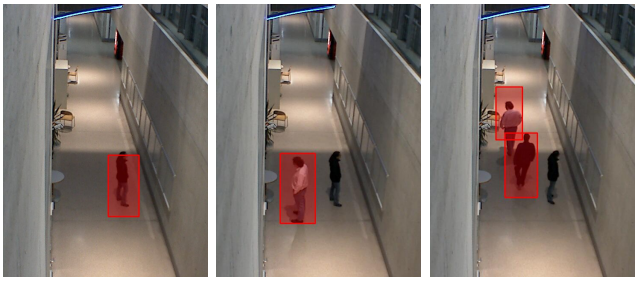
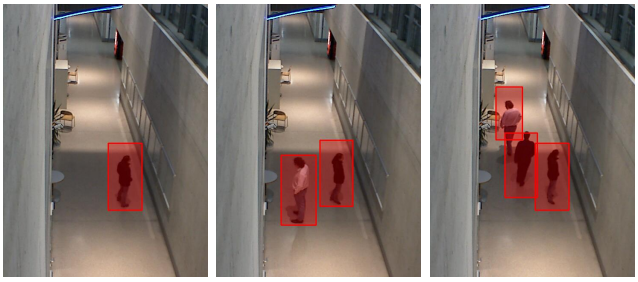



Figure 4. Recall-Precision Curves for the Corridor Sequence, containing objects that are not moving over a long period of time. The proposed approach clearly outperforms the CG approach, which yields a worse recall caused by short-term-drifting.



Frame 99 Frame 164 Frame 274

(a) Classifier Grid [5]



Frame 99 Frame 164 Frame 274

(b) Proposed approach

Figure 5. 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.

## V. Conclusion

We presented a method for real-time object detection for stationary cameras based on classifier grids. Our approach aims to solve the problem of short-term drifting, arising from fixed update strategies using the current input image to update the negative representation. Hence, 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 apply a multiple-instance learning (MIL) strategy. However, we had to modify the original multiple-instance learning idea (inverse MIL), since in our case the ambiguity lies in the negative examples. We collect a temporal bag of negative samples from the input images and adapted the on-line MILBoost [9] to fit to our problem. This new update strategy avoids the problem of short-term drifting for the classifier grid approach, which is clearly shown in the experiments.

## Acknowledgements

This work was supported by the FFG project HIMONI under the COMET programme in co-operation with FTW, the FFG project SECRECT under the Austrian Security Research Programme KIRAS, and the Austrian Science Fund (FWF) under the doctoral program Confluence of Vision and Graphics W1209.

## References

[1] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *Intern. Journal of Computer Vision*, vol. 77, no. 1–3, pp. 259–289, 2008.

[2] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. I, 2005, pp. 886–893.

[4] H. Grabner, P. M. Roth, and H. Bischof, "Is pedestrian detection really a hard task?" in *Proc. IEEE Workshop on PETS*, 2007.

[5] P. M. Roth, S. Sternig, H. Grabner, and H. Bischof, "Classifier grids for robust adaptive object detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[6] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. I, 2006, pp. 260–267.

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[8] P. A. Viola, J. C. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Advances in Neural Information Processing Systems*, 2005.

[9] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online mulitple instance learning," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[10] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004.