Learning of Scene-Specific Object Detectors by Classifier Co-Grids

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

Abstract

Recently, classifier grids have shown to be a considerable alternative to sliding window approaches for object detection from static cameras. The main drawback of such methods is that they are biased by the initial model. In fact, the classifiers can be adapted to changing environmental conditions but due to conservative updates no new object-specific information is acquired. Thus, the goal of this work is to increase the recall of scene-specific classifiers while preserving their accuracy and speed. In particular, we introduce a co-training strategy for classifier grids using a robust on-line learner. Thus, the robustness is preserved while the recall can be increased. The co-training strategy robustly provides negative as well as positive updates. In addition, the number of negative updates can be drastically reduced, which additionally speeds up the system. In the experimental results these benefits are demonstrated on different publicly available surveillance benchmark data sets.

1. Introduction

The first step in many visual surveillance applications is to identify the objects-of-interest (object detection). The most prominent approach is to apply a sliding window technique (*e.g.*, [5,6]), where each image patch is tested whether it is consistent with a previously learned model or not and all consistent patches are reported. Even though for most surveillance applications 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.* [9,15,16,22]). Hoiem *et al.* [9] proposed to use context information to reduce the number of false positives when applying a fixed generic object detector. In contrast, other approaches directly estimated a scene-specific object detector by incorporating information from the observed scenario via on-line learning [15, 16, 22].

Even more specific and thus more efficient classifiers can be trained using classifier grids (*e.g.*, [8, 17, 18]). 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, which, in turn, allows to apply less complex classifiers speeding up the detection procedure. A more detailed overview and discussion is given in Section 2.1.

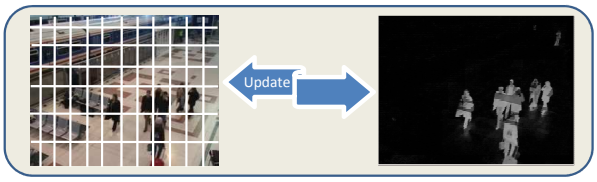

Figure 1. Co-grid classification: the grid classifiers on the left side are updated using labels generated by a second independent co-trained classifier evaluated on the background subtracted image.

By using on-line classifiers within the classifier grid the system is able to adapt to changes in the scene. Hence, those variabilities have not to be handled by the classifier and even less complex models can be applied. However, on-line learning requires to (robustly) include unlabeled data. Typical approaches to incorporate unlabeled data are semi-supervised learning (*e.g.* [7]), self- or co-training (*e.g.*, [4, 12]), or the application of oracles[1] (*e.g.*, [15, 21]). All of these methods explore unlabeled data to gain new information. 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.

The goal of this work is to develop an adaptive object detection system allowing to robustly include new scene specific samples while preserving the accuracy. In particular, this is realized by extending the ideas of classifier-grids by a robust "orthogonal" label generator (oracle) and the appli-

[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.

cation of a more suitable (*i.e.*, multi-class) learning method. This is illustrated in Figure 1. The oracle is initially trained by co-training following the ideas of Levin *et al*. [11]. However, to avoid drifting, after an initialization phase the co-training is stopped. In fact, it can be shown that the applied representation is invariant to most environmental changes and thus such a classifier provides excellent positive as well as negative updates. Moreover, the number of required updates is drastically reduced, speeding up the whole system. In the experiments, we demonstrate our approach on different publicly available benchmark data sets and show that compared to existing approaches including scene specific information helps to increase the recall while preserving the accuracy.

The rest of the paper is structured as follows. First, in Section 2 we review the ideas of classifier grids and co-training, which build the basis for our approach. Next, we introduce our new co-grid approach in Section 3 and give an experimental evaluation on publicly available standard benchmark data sets in Section 4. Finally, we summarize and conclude the paper in Section 5.

2. Classifier Grids and Co-Training

Stationary camera setups enable training of scene specific classifiers. Thus, compared to general detection tasks (*e.g.*, [5,6]), where a classifier has to work for any given scenario, the complexity of the classification task is drastically reduced, and more efficient and more compact classifiers can be applied. Moreover, by using temporal information and an on-line learning algorithm (*e.g.*, [15]) the classifier is able to adapt to changing environments, which further reduces the classifiers' complexity. This, however, requires that new unlabeled data is considered during the update process. In the following, we discuss the two scene-specific learning approaches that are relevant for this work.

2.1. Classifier Grids

The main concept of classifier grids [8] 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 2. 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 exploit information from labeled and unlabeled data, Grabner *et al*. [8] used fixed update rules. The positive updates are performed on a (small) labeled set whereas the negative samples are extracted from the underlying image patch described by one grid-classifier. These fixed rules have three main disadvantages: first, wrong negative updates may occur; second, even if the positive information is kept fix, positive updates are required; third, no new positive information is acquired.

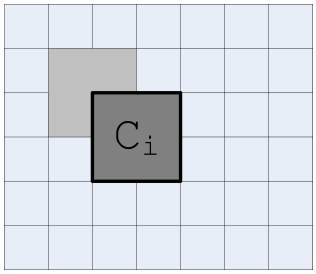

Figure 2. Grid-based classification: a highly overlapping grid of classifiers is placed over the image.

The first two problems can be reduced/avoided by the combination of two generative models as proposed by Roth *et al*. [17]. In particular, they apply a pre-trained generative model for the positive class and an adaptive generative model for the negative class, which is updated using samples from the scene. In this way the strong positive prior inhibits fast temporal drifting while ensuring the required adaptivity. Moreover, since the positive model is kept fix, the number of required updates is reduced. 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). This can easily be prevented by applying an additional background model to generate the data for the negative updates. However, the model for the foreground guarantees that the system recovers quickly (within a few frames), which allows to use such a setup for longterm tasks. Thus, when adapting over time the initial performance can be assured even if the system is running for a longer time,

The third problem, not gaining new positive information during learning, was addressed by Stalder *et al*. [18] by introducing context-based grid detectors to extract additional positive information from a specific scene. For that purpose they introduced three different ways to generate new scene specific samples: a fixed detector, a tracker, and 3D-context information from multiple camera views (if available). The negative samples are still generated from the scene (*i.e.*, extracted from a background model or patches showing identified static objects). In this way the recall can be drastically increased, but at the expense of the precision.

2.2. Co-Training

A different way to train scene-specific classifiers by exploiting information of unlabeled data from a scene was proposed by Levin *et al*. [11]. In particular, starting with a small number of hand-labeled samples they generated additional labeled examples by applying co-training [4] of two boosted off-line classifiers. One is trained directly from gray-value images whereas the other is trained from background subtracted images. The additional labels are generated based on confidence-rated predictions. Using the addi-

tionally labeled samples the training process is started again from scratch. In this way better classifiers can be obtained.

Co-training [4], in general, exploits the redundancy of unlabeled input data. The main idea is to train two initial classifiers h_1 and h_2 on some labeled data $\mathcal{D}^L$ and then let these classifiers update each other using unlabeled data $\mathcal{D}^U$. An update is performed if one classifier is confident on a sample whereas the other one is not. Since Abney [1] showed that co-training classifiers aim to minimize the error on the labeled samples while increasing the agreement on the unlabeled data, it is clear that the unlabeled data can help to improve the margin of the classifiers and to decrease the generalization error. Thus, co-training has recently become popular in the field of computer vision and was applied for a variety of applications including background modeling [23], learning an object detector [11], or tracking [13]

Since the approach of Levin *et al*. [11] is based on off-line classifiers, it is not suitable for an adaptive real-world detection system. However, since on-line boosting has become popular for visual learning (*e.g*., [2, 10]), having an initial classifier of sufficient accuracy the off-line classifier can easily be replaced by an on-line method still preserving the required properties. In fact, in Liu *et al*. [13] a proof for error bounds for on-line boosting in co-training was given. Moreover, the originally strong condition of conditionally independent classifiers was later relaxed by several authors (*e.g*., [1, 3]). Wang and Zhou [20] provided a PAC-style proof that co-training can converge to good accuracy if the classifiers are strong and highly uncorrelated. Thus, co-training can typically also be applied if, in principle, the learners are strong and low-correlated.

3. Classifier Co-Grids

The main disadvantage of existing grid-based object detectors, as discussed in Section 2.1, is that they are either biased by an off-line estimated prior or that adding new scene specific samples degrades the precision. However, the goal would be to increase the recall by adding scene specific samples, but keeping the classifiers' accuracy on the given level. In the following, we introduce an extended update scheme that assures both, a higher recall and stability.

Our main concept is depicted in Figure 1. The key idea is to use an independent "orthogonal" information source to provide stable positive and negative updates from the scene. To get such an "orthogonal" information source we adopted the visual co-training approach of Levin *et al*. [11]. In fact, we also apply background subtraction (*i.e*., approximated median background model [14]) to exploit the information given by this additional view on the data. In contrast to Levin *et al*., co-training is performed only in an initial phase. Later on this classifier is kept fix and used as an oracle for two reasons. First, not all situations can be handled by co-training in a robust manner. Hence, if too many wrong updates are performed the co-trained classifier would start to drift and finally fails totally. Second, as illustrated in Figure 3, most environmental changes are eliminated by the background subtraction and the variability in the positive class vanishes. Thus, no further information can be gained.

Figure 3. Different illumination conditions with the corresponding background subtracted images. Even in case of totally different illumination conditions and differently appearing objects the background subtracted image gives a similar representation of the object.

3.1. Co-Training Stage

During the initial stage our system is trained in a co-training manner. Given n grid classifiers G_j operating on gray level image patches $\mathbf{X}_j$ and one compact classifier C operating in a sliding window manner on background subtracted images $\mathbf{B}$. To start co-training, the classifiers G_j as well as the classifier C are initialized with the same off-line pre-trained classifier, then the classifiers co-train each other. A confident classification (positive and negative) of a classifier G_j is used to update the classifier C with the background subtracted representation at position j. Vice versa, a confident classification of classifier C at position j generates an update for classifier G_j. Due to the off-line pre-trained prior, already capturing the generic information, only a small number of updates is sufficient to adapt the classifiers to a new scene or changing environmental conditions. The update procedure during the initialization for a specific grid element j is summarized in Algorithm 1.

3.2. Detection Stage

After the initial stage, as described above, the classifier C is not longer updated and is applied as an oracle to generate new positive and negative samples. Hence, we can abstain from fixed update rules, which are broadly used for classifier grids. Moreover, we perform negative updates for the classifiers G_j only if they are necessary, *i.e*., if the scene is changing. Even if the oracle classifier C has a low recall, the precision is very high. Thus, only very valuable updates are generated, increasing the performance of the classifiers G_j. In particular, a confident classification result of classifier C at position j generates an update for all classifiers G_i,

$i = 1, \ldots, n$. In this way new scene specific positive samples are disseminated over the whole classifier grid. Negative updates are performed for classifiers G_j if there is no corresponding detection reported at this position for classifier C. The update procedure during the detection phase for a specific grid element j is summarized in Algorithm 2.

Algorithm 1 Co-Grid Initialization

Input: grid-classifier G_j^{t-1}, classifier C^{t-1}
Input: grid-element $\mathbf{X}_j$, BGS patch $\mathbf{B}_j$

if $C^{t-1}(\mathbf{B}_j) > \theta$ **then**
  $update\left(G_j^{t-1}, \mathbf{X}_j, +\right)$
else if $C^{t-1}(\mathbf{B}_j) < -\theta$ **then**
  $update\left(G_j^{t-1}, \mathbf{X}_j, -\right)$
end if

if $G_j^{t-1}(\mathbf{X}_j) > \theta$ **then**
  $update\left(C^{t-1}, \mathbf{B}_j, +\right)$
else if $G_j^{t-1}(\mathbf{X}_j) < -\theta$ **then**
  $update\left(C_{t-1}, \mathbf{B}_j, -\right)$
end if

Output: grid-classifier G_j^t, classifier C^t

Algorithm 2 Co-Grid Update

Input: grid-classifier G_j^{t-1}, classifier C
Input: grid-element $\mathbf{X}_j$, BGS patch $\mathbf{B}_j$

if $C(\mathbf{B}_j) > \theta$ **then**
  $\forall i : update\left(G_i^{t-1}, \mathbf{X}_j, +\right)$
end if

if $C(\mathbf{B}_j) < -\theta$ **then**
  $update\left(C_j^{t-1}, \mathbf{X}_j, -\right)$
end if

Output: grid-classifier G_j^t

3.3. On-line Learning

In general, any on-line learner can be applied within the co-grid approach. However, to increase the robustness (the co-trained oracle may still suffer from a small amount of label noise), we apply on-line TransientBoost [19], which allows to combine reliable (labeled) data with unreliable data (unlabeled from the scene). Thus, robustly new positive information can be gained, especially increasing the recall but preserving the accuracy. The main idea of TransientBoost is to train a binary classifier which is based on an internal multi-class representation. Assuming a strong classifier $F_m(x) = \sum_{t=1}^{m} f_t(x)$, this is realized by using histograms as weak learners $f_t(x)$. Moreover, histograms are highly suitable for a combined off-line/on-line learning since they can easily be estimated for both domains. In our case, we apply three classes: one for the reliable positive data (+1), one for the possibly unreliable positive data (+2), and one for the classifier-specific background (-1). The class +1 is trained off-line and is kept fix whereas the others are updated on-line over time. To finally get a binary classification, compared to multi-class methods the weight update during training and the evaluation have to be adapted. In our case the weight w_n for the current sample x is estimated as $w_n = -\ell'\left(sign(y_t)F_m(x)\right)$, where y_t is the computed label, F_m is the strong classifier, and ℓ is the loss function. In addition, the evaluation function is set to $sign\left(F_m(x)\right)$. For more details we refer to [19].

4. Experimental Results

In the following, we demonstrate our approach for two different tasks, *i.e.*, pedestrian detection and car detection. We evaluated our approach on two publicly available standard benchmark data sets for pedestrian detection, *i.e.*, the *PETS 2006* data set[2] and the *PETS 2009* data set[3] and one publicly available data set for car detection, *i.e.*, *AVSS 2007*[4]. In particular, we want to show that incorporating scene specific information can help to increase the recall while using the proposed approach the accuracy can be preserved. In addition, we compared our approach for all data sets to the adaptive grid-based object detector of Roth (GB) *et al*. [17] and a fixed generic state-of-the art detector, namely the deformable part model of Felzenszwalb *et al*.[5](FS) [6]. Both can be applied for person as well as for car detection. The latter one does not use any scene specific specific knowledge, however, to enable a fair evaluation similar to Hoiem *et al*. [9] all false positives that are not on the given ground-plane are removed.

4.1. PETS 2006

First of all, we run the experiments on the *PETS 2006* data set showing the concourse of a train station, where we have chosen a representative sequence consisting of 308 frames (720x576 pixels). The corresponding recall-precision curves (RPC) are shown in Figure 4. It clearly can be seen that the fixed generic detector as well as existing grid-based approach can be outperformed in terms of both, recall and precision (*i.e.*, for a comparable precision the recall can be increased from 0.45 to 0.80). This clearly shows that including scene-specific data is highly beneficial. Moreover, illustrative results for this sequence are shown in Figure 5.

[2] http://www.pets2006.net
[3] http://www.pets2009.org
[4] http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html
[5] http://people.cs.uchicago.edu/~pff/latent

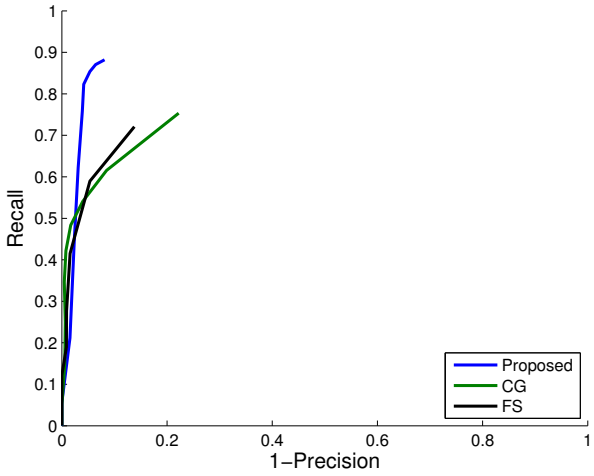

Figure 4. RPCs for the *PETS 2006 Sequence*.

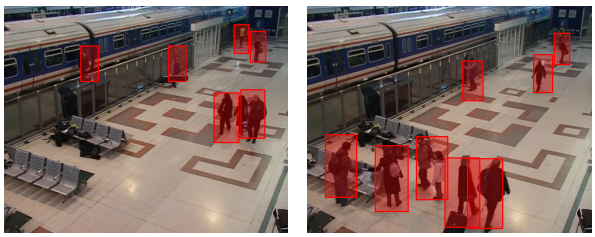

Figure 5. Illustrative detection results of our approach for the *PETS 2006* Sequence.

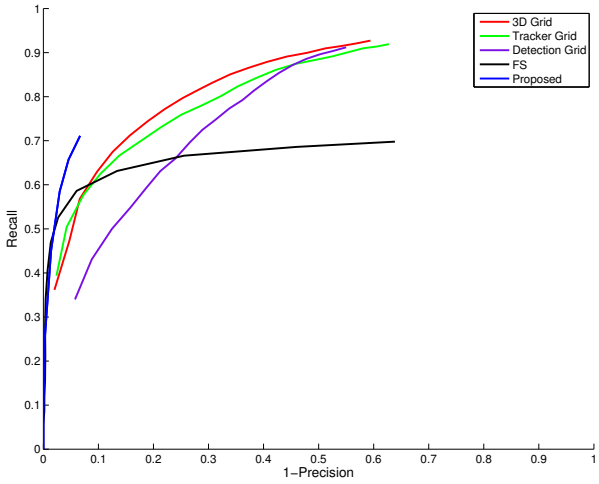

Figure 6. RPCs for the *PETS 2009 Sequence*.

Figure 7. Illustrative detection results of our approach for the *PETS 2009* Sequence.

4.2. PETS 2009

Next, we evaluated our approach on the *PETS 2009* data set. The selected sequence consisting of 188 frames (768x576 pixels), which is very highly crowded and thus very challenging, is the same that was used in Stalder *et al*. [18]. Thus, for this test set we additionally give a comparison to three different grid-based approaches proposed by Stalder *et al*. The obtained RPCs are shown in Figure 6. Again it can be seen that the proposed approach allows to robustly include new information increasing the recall but not tackling the precision. Moreover, it can be seen that even though Stalder *et al*. [18] use more sophisticated methods to incorporate new information (*e.g*., tracking or 3D information) our approach yields a significantly better precision on the same recall level. Illustrative detection results of our approach on this data set are shown in Figure 7.

4.3. AVSS 2007

Finally, to show that the proposed method is not limited to person detection, we demonstrate it for car detection. In particular, we run the experiments on the AVSS 2007 data set, where we evaluated on the first 500 frames (720x576 pixels) of the vehicle detection sequence AVSS_PV_Hard. The RPCs for the proposed approach, CG, and FS are shown in Figure 8. It can be seen that the FS approach clearly can be outperformed and that compared to the CG approach, even given an excellent baseline, the recall can be further increased. Again illustrative detection results of our approach are shown in Figure 9.

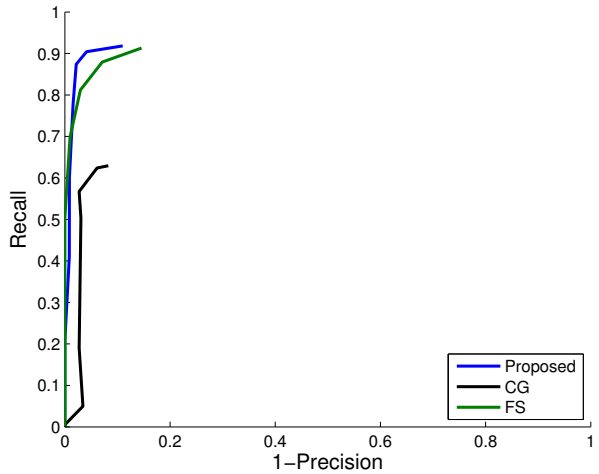

Figure 8. RPCs for the *AVSS 2007 Sequence*.

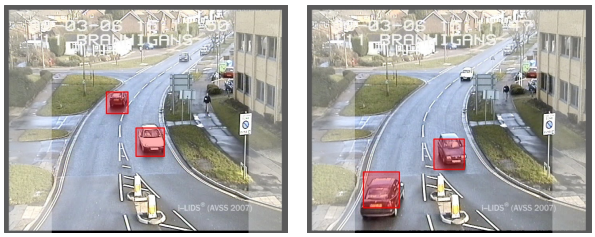

Figure 9. Illustrative detection results of our approach for the *AVSS 2007* sequence (detection results within the fully colored region).

5. Conclusion

We presented a robust real-time object detection system for stationary cameras which is able to adapt to a scene without drifting. The approach is based on the idea of classifier grids, *i.e.*, each image location is represented by one separate classifier. Since existing systems are either biased by a prior or the accuracy is decreased when performing positive updates the goal of this work is to benefit from positive scene specific information while preserving the accuracy. This is realized by the combination of an oracle, which is initialized using co-training, and a robust on-line learner. In general, any on-line learning method can be applied, but in particular to increase the robustness, we apply TransientBoost, an on-line multi-class boosting method allowing for preserving the prior information. We demonstrated our system for three different publicly available data sets (person and car detection), where we show that the recall can be increased while preserving the accuracy and that existing methods can be outperformed.

Acknowledgment

The work was supported by the FFG project HIMONI under the COMET programme in co-operation with FTW and the FFG project SECRET (821690) under the Austrian Security Research Programme KIRAS.

References

[1] S. Abney. Bootstrapping. In *Proc. Annual Meeting of the Association for Computational Linguistics*, pages 360–367, 2002.

[2] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online mulitple instance learning. In *Proc. CVPR*, 2009.

[3] M.-F. Balcan, A. Blum, and K. Yang. Co-training and expansion: Towards bridging theory and practice. In *NIPS*, pages 89–96, 2004.

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

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

[6] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. CVPR*, 2008.

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

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

[9] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *Proc. CVPR*, volume II, pages 2137–2144, 2006.

[10] C. Leistner, A. Saffari A. A., P. M. Roth, and H. Bischof. On robustness of on-line boosting - a competitive study. In *Proc. IEEE On-line Learning for Computer Vision Workshop*, 2009.

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

[12] L.-J. Li, G. Wang, and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. In *Proc. CVPR*, pages 1–8, 2007.

[13] R. Liu, J. Cheng, and H. Lu. A robust boosting tracker with minimum error bound in a co-training framework. In *Proc. ICCV*, 2009.

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

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

[16] P. M. Roth and H. Bischof. *Machine Learning Techniques for Multimedia*, chapter Conservative Learning for Object Detectors, pages 139–158. Springer, 2008.

[17] P. M. Roth, S. Sternig, H. Grabner, and H. Bischof. Classifier grids for robust adaptive object detection. In *Proc. CVPR*, 2009.

[18] S. Stalder, H. Grabner, and L. v. Gool. Exploring context to learn scene specific object detectors. In *Proc. IEEE Workshop on PETS*, 2009.

[19] S. Sternig, M. Godec, P. M. Roth, and H. Bischof. Transientboost: On-line boosting with transient data. In *Proc. IEEE Online Learning for Computer Vision Workshop (CVPR)*, 2010.

[20] W. Wang and Z.-H. Zhou. Analyzing co-training style algorithms. In *Proc. European Conf. on Machine Learning*, pages 454–465, 2007.

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

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

[23] Q. Zhu, S. Avidan, and K.-T. Cheng. Learning a sparse, corner-based representation for background modelling. In *Proc. ICCV*, volume I, pages 678–685, 2005.