Robust Person Detection by Classifier Cubes and Local Verification *

Sabine Sternig, Hayko Riemenschneider, Peter M. Roth, Michael Donoser, and Horst Bischof

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

Abstract
Classifier grids have shown to be an alternative to sliding window approaches for object detection from static cameras. However, existing approaches neglected two essential points: (a) temporal information is not used and (b) a standard non-maxima suppression is applied as post-processing step. Thus, the contribution of this paper is twofold. First, we introduce classifier cubes, which exploit the available temporal information within a classifier grid by adapting the local detection likelihood based on preceded detections. Second, we introduce a more sophisticated post-processing step to verify detection hypotheses by comparing a local figure/ground segmentation to a provided prototype model. Experiments on publicly available data demonstrate that both extensions improve the detection performance.

1. Introduction

Object detection is of high interest in various applications in computer vision. The most prominent approach is to use a sliding window technique, which scans over all possible locations in the image and subsequently evaluates a pre-trained classifier. Since in this way overlapping detections are generated, local maxima of the classifier responses are considered as object location hypotheses. When having a stationary camera, which is a reasonable assumption for many surveillance scenarios, classifier grids [4, 12, 13] have shown to be a viable alternative. 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 since each classifier only has to discriminate the object-of-interest from the background at one specific image location.

Recent work on classifier grids has mainly focused on stability and including new scene specific updates. To avoid drifting, either fixed update strategies [4] or a strong positive prior [12] were introduced. In this way a stable system over time can be ensured, however, since the models are strongly biased by the prior no new positive information is gained and the recall cannot be increased. In contrast, in [13] context-based learning was used to gain new scene specific information allowing to increase the recall, however, decreasing the precision at the same time. In general, all of these methods neglect two essential points: (a) they operate on single images only, even though temporal information

*This work was supported by the FFG project HIMONI under the COMET programme in co-operation with FTW, the FFG project CityFit (815971/14472-GLE/ROD) under the FIT-IT programme, by the FFG project SECRECT (821690) under the Austrian Security Research programme KIRAS, and the Austrian Science Fund (FWF) under the doctoral programme Confluence of Vision and Graphics W1209.

such as given by a video stream is available and (b) each classifier has its own characteristics and therefore standard post-processing approaches are often not meaningful.

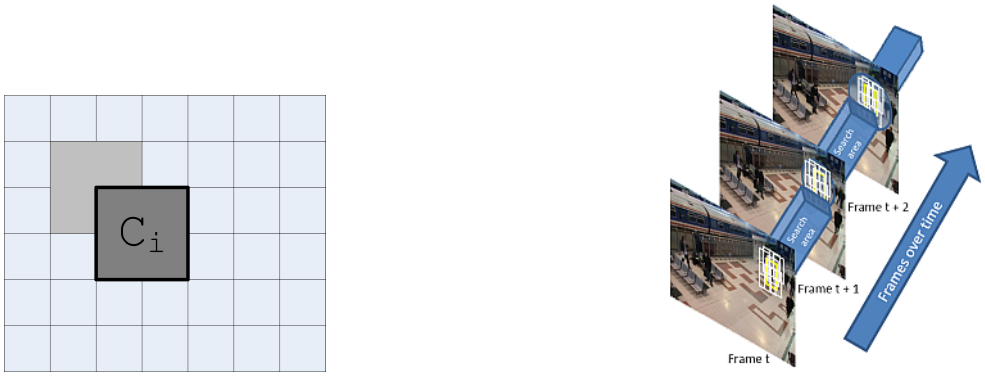
The goal of this work is to exploit the power of classifier grids while avoiding the shortcomings mentioned above. Thus, our contributions are twofold. First, in Section 2 we describe how the idea of classifier grids can be extended by a temporal dimension obtaining classifier cubes. In fact, incorporating the available temporal information allows to increase the recall without degrading the precision or the stability. Second, in Section 3 we introduce a shape-based post-processing approach built on Maximally Stable Extremal Region (MSER) [8] detection to verify provided object hypotheses. By analyzing the accuracy of the segmentations and generating a corresponding score, in a more sophisticated way more suitable candidates can be selected. In Section 4 the benefits of both contributions are demonstrated in the context of pedestrian detection.

2. Classifier Cubes

The main concept of classifier grids [4] is to sample an input image by using a highly overlapping grid, where each grid element $i = 1, \ldots, N$ corresponds to one specific classifier c_i. This is illustrated in Figure 1(a). 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 [4, 12]. 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 [3] has proven to be a considerable trade-off between speed and accuracy.

To incorporate the local spatial information, for existing methods [4, 12, 13] the classifiers within a classifier grid are updated all the time. However, for the detection the temporal information is not used at all. In the following we take advantage of the temporal constraints by transferring (detection) information to succeeding frames. This idea is related to tracking by detection, which has become popular over the last years (e.g., [1, 3, 16]). These approaches use the detection result of one frame in order to find the object of interest in the next frame. Since the object-of-interest is expected to move only within a restricted area between two successive frames, the search region can be restricted. In contrast, in this work we are not interested in the trajectories of the objects-of-interest. However, by adding a temporal dimension to the classifier grids we want to improve the detection results by introducing *classifier cubes* as illustrated in Figure 1(b).

Formally, given a classifier $c_{i,t}$ at time t with center position (x_i, y_i), the classifier cube C_i is defined as follows: $C_i = \{\{c_{j,t-1}\, c_{j,t}\, c_{j,t+1}\} \mid d\left((x_i, y_i), (x_j, y_j)\right) < \theta\}$, where θ defines the local spatial neighborhood around classifier $c_{i,t}$ where a detection was reported at time t. In other words, a classifier cube C_i consists of all classifiers from frame $t-1$ to frame $t+1$ which are within a defined local spatial neighborhood around classifier $c_{i,t}$. This local and temporal structure described by the classifier cube $C_{i,t}$ can now be used to reduce the detection sensitivity s_j. By increasing the sensitivity, objects close to the decision boundary, which were missed otherwise, can now be detected resulting in an increased recall. However, since this step is only performed in a very local neighborhood of expected true positives the precision stays the same. The update process is not influenced by this evaluation strategy and any on-line learner can be applied. Please note, that the evaluation is still performed on a single-frame basis – the temporal information is only incorporated by modifying the sensitivity.



(a) (b)

Figure 1. (a) Classifier grids and (b) classifier cubes using temporal information.

Instead of propagating the detections over only one foregoing and one subsequent frame, it is also possible to propagate the detection information over a number of frames n over time. In this case the movements of the objects have to be considered, and thus the search location (i. e., the radius of classifier cube) is increased towards the end of the cube.

As shown in the experiments in Section 4, this extension allows to improve the overall detection performance, i. e., increasing the recall, while preserving a high recall due to the fixed update strategies. Even in case of a false positive, which results in increasing the probability of a detection in the next frame, this false positive is not propagated over time. Since the background is updated all the time, the false positives can efficiently be removed.

3. Verification by Figure/Ground Segmentation

In general, segmentation is used as a tool to qualitatively improve the performance of object detection systems [10, 14] in three forms. First, detection results are augmented with a segmentation of the accurate object outline. For example, Wang et al. [15] use shape context to hypothesize detection locations and apply a modified normalized cut to provide a figure/ground segmentation. Zhao et al. [17] use Chamfer matching between edge templates to obtain hypotheses. Then they use an interactive kernel density estimation of the two Gaussian distributions for foreground and background to obtain a stable segmentation. Second, detection hypotheses are verified by a local segmentation to discard false positives. In this line Rihan et al. [11] combine a face detector with a posterior probability distribution to guide a real-time Markov Random Field (MRF) framework. By including energy terms for shape, contrast, and color they can improve the segmentation obtained by a graphcut. Ramanan [10] uses learned Gaussian distributions in a graphcut segmentation to extract local figure/ground separations. The binary segmentations are then learned in a support vector machine (SVM) classifier to provide the final verification of the detector hypotheses. Third, tight bounding box constraints can be used to guide even better segmentation results. Recent work by Lempitsky et al. [6] uses a pinpointing algorithm to obtain a foreground segmentation which tightly fits such a provided bounding box.

The limitations of these methods are mainly the quality of the extracted segmentations and the level of supervision required to obtain a well-fitting figure/ground separation at reasonable runtime. Our goal is also to verify the obtained object hypotheses, however, we compare a figure/ground segmentation of the provided bounding box to an object prototype model, i.e., the mean shape of a training dataset. The

resulting similarity score from the comparison is used to verify the detection hypotheses. Our findings show that the foreground segmentation corresponds directly to the mean prototype segmentation if the underlying object is present, see Figure 2. In this example multiple objects are present and the figure segmentation delivers foreground blobs for all of them. The similarity evaluation is shown as a heat map on top of the image. The two graphs on the right show multiple similarity responses shifted along the x and y axis. The scores indicate how the true location of an object is located at the peaks of the similarity evaluation.

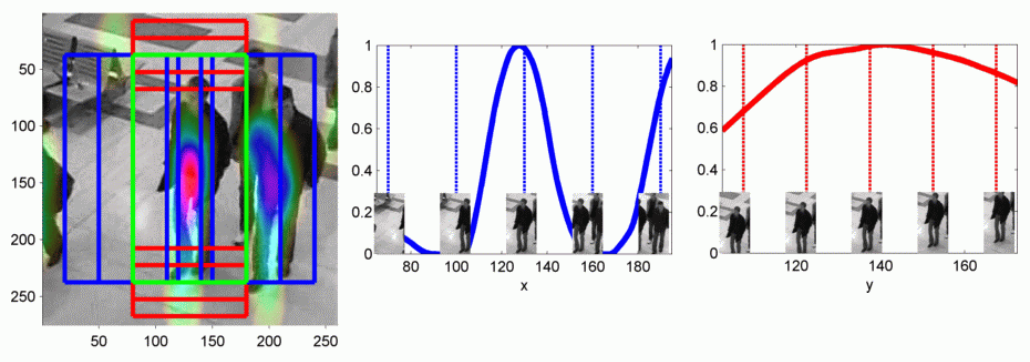


Figure 2. Intersection over union (IOU) similarity scores for multiple detections around the true person location delivered by a detector.

In our work, we define any centered blob-like object as foreground which represents a stable separation from the background. To get a stable figure/ground separation, the main idea is to detect Maximally Stable Extremal Regions (MSER) [8], which have originally been introduced as interest region detectors for wide-baseline stereo matching. In general, MSERs are stable extremal regions, which can be considered the connected regions of threshold operations of the image. By using a stability criterion a region is compared to the same region in other intensity levels and finally those regions are selected whose size is stable relative to the changes in intensity levels. The advantages of this segmentation method are analogous to those of the interest point detector, which has proven to be one of the most repeatable detectors. It provides invariance to scale and photometric changes and it is also covariant to adjacency preserving transformations. The MSER-based figure / ground segmentation thus extracts repeatable foreground blobs, which are well-suited for the desired separation of foreground and background.

Our verification procedure consists of the following steps. First, MSERs are detected within the scaled (fixed height and width) bounding box, located at an object hypothesis. This delivers blob-like regions which correspond to foreground objects. Second, the obtained figure/ground segmentation is compared to a mean binary prototype shape. This shape is obtained by calculating the mean over all extracted MSERs in a set of positive training images. The similarity between two binary segmentations is calculated as the pixel-wise intersection over the union (IOU). See Section 4 for details on the similarity score. As shown in the experiments in Section 4, the proposed approach delivers competitive segmentation results for the object-of-interest at greatly reduced runtime. Both benefits prove to be valuable properties for any post-processing using segmentation for hypotheses verification.

4. Experiments

In the following, we demonstrate the benefits of the proposed person detection system. In particular, the experiments are split into two parts. First, we analyze the segmentation-based post-processing

by comparing different segmentation methods outlining that our proposed MSER-based verification provides the best performance. Second, we demonstrate the overall system on a publicly available benchmark data set for person detection showing both, the benefits of adding temporal information (classifier cubes) and using the introduced post-processing step.

4.1. Segmentation Based Verification

First of all, we give a qualitative and quantitative comparison of our MSER-based figure extraction to a variety other figure / ground separation methods: (a) a multi-scale normalized cut method (*ncut*) [2], where the image patch is separated into multiple segments (n=5) and the center segment is selected as the foreground; (b) a levelset method [7], which uses either a mean binary shape (*ls*) or (c) a centered rectangular region as initialization (*ls_box*); (d) a graph cut method given local Gaussian distributions of foreground and background pixels (*gc*) and (e) additionally using the mean shape as prior (*gc_p*); (f) extracting Canny edges and comparing them to the boundaries of the mean binary shape (*canny*). For each of these methods we use normalized cross-correlation between the figure/ground segmentation and a ground truth binary mask as similarity measure. Finally, we also analyze the similarity using a standard normalized cross-correlation between the raw image patch and the mean appearance image.

A qualitative analysis of figure/ground segmentations is given in Figure 3. In the first row, segmentation results for a true positive detection hypothesis are shown, whereas the second row shows results for a false positive containing background and no object instance. The examples in Figure 3 show that almost all methods are able to extract well-fitting object segmentations. The quality of the segmentations achieved by methods using shape priors (*ls*, *ls_box* and *gc_p*) is superior to the unsupervised methods. However, the opposite effect can be recognized from the second row. The methods using shape priors hallucinate objects in the background image patches. The unsupervised methods extract unconstrained figure segmentations, with a significantly different shape compared to the prototype model.

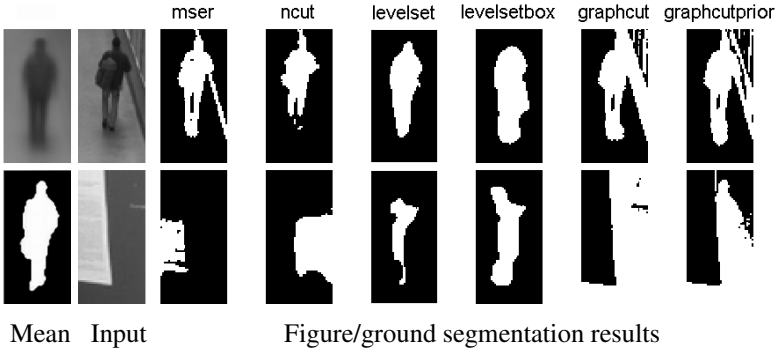


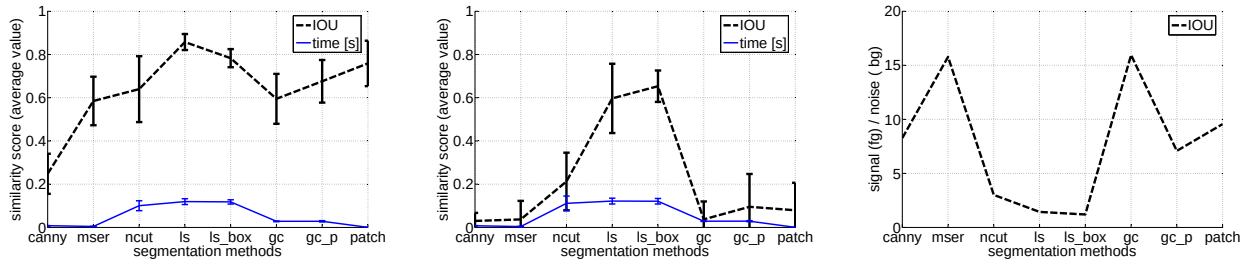
Figure 3. Results for figure/ground segmentation for patches with an object (top) and background without an object (bottom) using MSER, normalized cut, level sets, and graph cut.

For the quantitative comparison, we evaluate all methods on a benchmark dataset and compare it to a mean binary segmentation of the object. The dataset consists of 381 positive and 6944 negative images. For each image all segmentation methods are applied and a similarity score is determined. The similarity score is calculated by a normalized pixel-wise intersection over union, denoted as *iou*.

In particular, we evaluate the performance of the normalized cross correlation w.r.t. its ability to separate the foreground objects (persons) from random background patches. For that purpose, Figures 4 (a) and (b) illustrate the similarity scores of the comparison to the ground truth segmentations

for the true positive and true negative patches, respectively. Figure 4 (a) confirms the accurate segmentations provided by methods using a shape prior in the positive image patches. However, as shown in Figure 4 (b), these approaches also hallucinate figure objects in the negative image patches as can be seen by their higher similarity scores to the mean segmentation. The most important scores are shown in Figure 4 (c), illustrating the ratio between the true and false positive similarities (which should be as high as possible). As can be seen, our proposed MSER-based figure extraction performs very similar to the supervised graph cut segmentation, but at much lower computational costs. As a baseline we additionally show results without segmentation using a standard normalized cross-correlation (NCC) between the mean image and appearance of the image patches. This NCC also performs well on the positive images, however, it has a higher undesirable similarity for the negative images due to the relative high amount of background in the patches.

Extensive parameter studies show that two effective parameters for MSER extraction (intensity level delta and minimum size) have only minor effects on the similarity score. The minimum size only effects how large the figure segment must be and the similarity drops after we reach 50% (400px) of the actual foreground object (800px). Increasing the intensity level delta results in a similar mean similarity score, however, the standard deviation is increased. Thus we select the best performance for each parameter, which is 2 for delta and 50 for the minimum size.



(a) Positive class (b) Negative class (c) Ratio shows separability

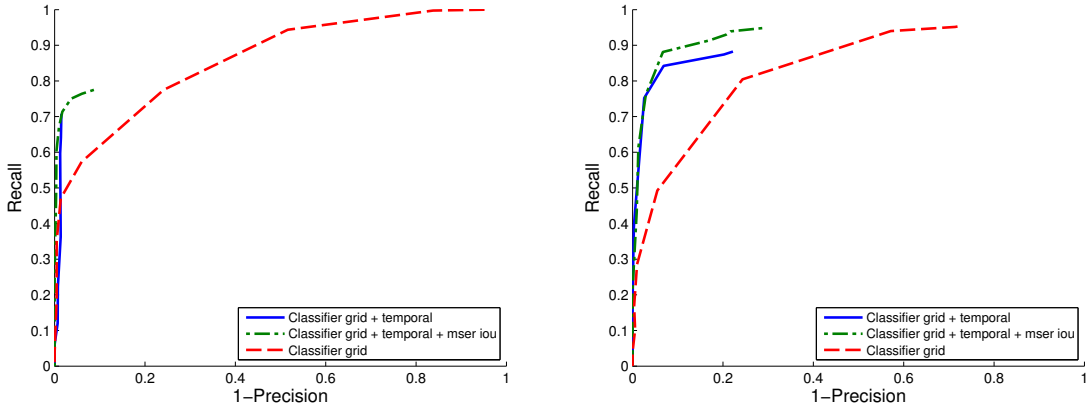
Figure 4. Performance (in similarity to mean binary shape) of figure/ground segmentation methods: (a) for positive (higher is better) and (b) negative (lower is better) training images. In addition, to illustrate the separability of the two classes, in (c) the similarity ratio of positive over negative samples is illustrated. MSER and graphcut perform best using a intersection over union (IOU) scoring on the binary segmentations.

4.2. Person Detection

Next, we demonstrate that both, the temporal classifier cubes and the segmentation based post-processing, clearly improve the classification, we run experiments on two different data sets. First, on the publicly available *PETS 2006* data set[1], showing a concourse of a train station consisting of 308 frames (containing 1714 pedestrians). Second, on a new data set, *corridor sequence*, with 900 frames (640x480), containing 2491 pedestrians, which we have generated in our lab. In general, the classifier cube method is quite general and not limited to a specific learner. However, to enable a fair comparison to the baseline approach of Roth et al. [12] (classifier grids), we also use an on-line boosting variant (On-line GradientBoost) [5] for the classifiers in the classifier cubes. These classifiers consist of 50 selectors, each of it containing 30 weak classifiers using Haar-like features. As proposed in [12], we pre-calculate the generative model for the object class off-line and continuously update the background class using an approximated median background model [9].

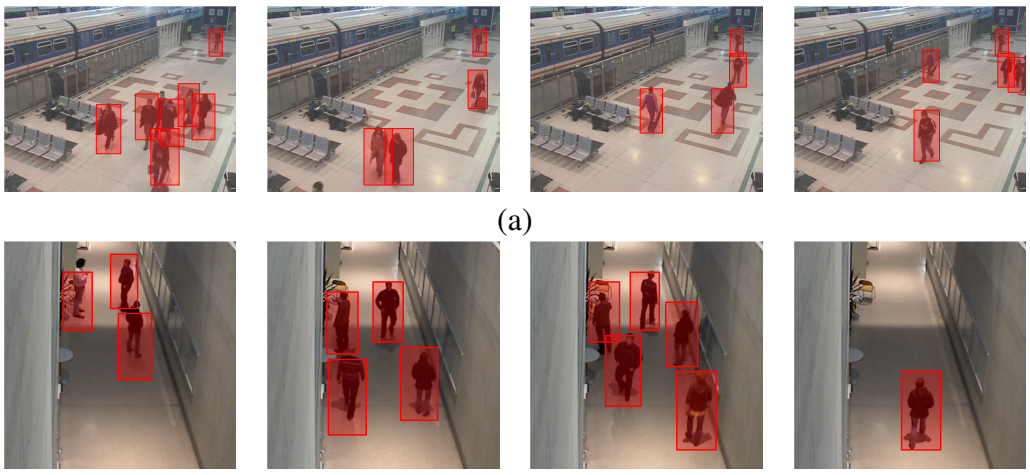
[1] http://www.pets2006.net

The obtained Recall-Precision Curves (RPC) for both test scenarios are given in Figure 5. It can be seen that by adding temporal information (classifier cubes) compared to the baseline approach the recall can clearly be increased. Moreover, using the more sophisticated segmentation based post-processing, where the IOU scores of the segmentation step are used as final detection confidence, further slightly increases the recall. Finally, in Figure 6 we show illustrative detection results of the combined approach.



(a) (b)

Figure 5. Recall Precision Curves for (a) PETS 2006 and (b) Corridor sequence.



(a)

(b)

Figure 6. Illustrative detection results for (a) PETS 2006 and (b) Corridor sequence.

5. Conclusion

In this work, we extended the idea of classifier grids, which have recently shown to be a considerable alternative to a sliding window technique for object detection from static cameras. In particular, our contribution was twofold. First, we extended the idea of classifier grids to a third dimension (classifier cubes), which, at a given level of precision, allows to increase the recall. Second, we introduced a novel post-processing step which uses a Maximally Stable Extremal Region based segmentation in order to validate the detection results provided by the classifier cubes. In the experimental evaluations we showed for the task of person detection that both extensions clearly improve the classification results compared to an existing grid classifier as well as to a static detector. In addition, we gave a detailed analysis of the MSER-based post-processing.

References

[1] S. Avidan. Ensemble tracking. In *Proc. Conf. on Computer Vision and Pattern Recognition*, 2005.

[2] T. Cour, F. Benezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. In *Proc. Conf. on Computer Vision and Pattern Recognition*, 2005.

[3] H. Grabner and H. Bischof. On-line boosting and vision. In *Proc. Conf. on Computer Vision and Pattern Recognition*, 2006.

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

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

[6] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp. Image segmentation with a bounding box prior. In *Proc. Intern. Conf. on Computer Vision*, 2009.

[7] C. Li, C. Xu, C. Gui, and M. Fox. Level set evolution without re-initialization: A new variational formulation. In *Proc. Conf. on Computer Vision and Pattern Recognition*, 2005.

[8] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. British Machine Vision Conf.*, 2002.

[9] N. McFarlane and C. Schofield. Segmentation and tracking of piglets. *MVA*, 1995.

[10] D. Ramanan. Using segmentation to verify object hypotheses. In *Proc. Conf. on Computer Vision and Pattern Recognition*, 2007.

[11] J. Rihan, P. Kohli, and P. Torr. Objcut for face detection. In *Proc. Indian Conf. on Computer Vision, Graphics and Image Processing*, 2006.

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

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

[14] Z. Tu, X. Chen nad A. Yuille, and S.-C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. In *Proc. Intern. Conf. on Computer Vision*, 2003.

[15] L. Wang, J. Shi, G. Song, and I. Shen. Object detection combining recognition and segmentation. In *Proc. Asian Conf. on Computer Vision*, 2007.

[16] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. In *IJCV*, 2007.

[17] L. Zhao and L. Davis. Closely coupled object detection and segmentation. In *Proc. Intern. Conf. on Computer Vision*, 2005.