IMPROVING PERSON DETECTION IN VIDEOS BY AUTOMATIC SCENE ADAPTATION

Roland Mörzinger and Marcus Thaler
Institute of Information Systems, Joanneum Research, Steyrergasse 17, 8010, Graz, Austria

Keywords: Object detection, Ground plane, Surveillance, Robust statistics, RANSAC, Adaptation.

Abstract: The task of object detection in videos can be improved by taking advantage of the continuity in the data stream, e.g. by object tracking. If tracking is not possible due to missing motion features, low frame rate, severe occlusions or rapid appearance changes, then a detector is typically applied in each frame of the video separately. In this case the run-time performance is impaired by exhaustively searching each frame at numerous locations and multiple scales. However, it is still possible to significantly improve the detector's performance if a static camera and a single planar ground plane can be assumed, which is the case in many surveillance scenarios. Our work addresses this issue by automatically adapting a detector to the specific yet unknown planar scene. In particular, during the adaptation phase robust statistics about few detections are used for estimating the appropriate scales of the detection windows at each location. Experiments with an existing person detector based on histograms of oriented gradients show that the scene adaptation leads to an improvement of both computational performance and detection accuracy. For scene specific person detection, changes to the implementation of the existing detector were made. The code is available for download. Results on benchmark datasets (9 videos from i-LIDS and PETS) demonstrate the applicability of our approach.

1 INTRODUCTION

The performance of computer vision applications can be optimized by incorporating scene context, such as the knowledge about background, ground plane and objects of interest. In the case of object detection, the task of object detection can be simplified by focusing only on the scales and image regions where the objects would typically appear. Consequently a considerable speedup and increase in accuracy can be achieved.

Prior work on exploiting scene context showed that object tracking can be improved by relying on the knowledge about a ground plane (Greenhill et al., 2008; Renno et al., 2002). This work estimates the depth of the scene at each pixel by observing moving objects in order to improve tracking of occluded moving regions. It builds on the valid assumption that in typical surveillance settings the object height in the image varies linearly with its vertical position in the image. The drawback of these approaches is that the linear model, i.e. the camera viewpoint consisting of gradient and horizon line, has to be defined manually. Another work (Hoiem et al., 2006) obtains an improvement over standard low-level detectors by putting objects in perspective and reasoning about the underlying 3D scene structure. Specifically, estimates about the rough scene surface geometry and the camera viewpoint supply likely scales of the objects in the image. These estimates were formed based on learning from a set of manually labeled horizons and available statistics for height distributions of 3D world objects. In their experiments they used the Dalal&Triggs person detector (Dalal and Triggs, 2005) to show how their approach improves object detection of pedestrians and cars. A framework for inferring scene information in monocular videos such as the relative depth and unevenness of ground is proposed in (Zhu et al., 2008). The occurrence probability of pedestrians at each location of the scene is learned in a semi-supervised fashion. This process requires a large amount of video data and a number of manually marked pedestrian samples which are collected over time at different positions in the scene. Recently, in (Stalder et al., 2009) the rough 3d scene context is explored for learning grid-based object detectors. This approach assumes overlapping calibrated views of the same scene so that corresponding regions from the different views can be used as training samples.

Möerzinger R. and Thaler M. (2010).
IMPROVING PERSON DETECTION IN VIDEOS BY AUTOMATIC SCENE ADAPTATION.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 333-338
DOI: 10.5220/0002820203330338

Recently a work that is able to deal with arbitrary ground surfaces using online learning has been proposed (Breitenstein et al., 2008). Multiple walkable surfaces of a scene are derived from the output of a pedestrian detector based on an entropy framework. According to the authors, this is the first work to exploit scene structure for optimizing the location-dependent scale range parameters used for improving object detection. They show that their method effectively limits the number of detection windows compared to an original pedestrian detector (Dalal and Triggs, 2005). Conceptually, the above work is most closely related to ours, but we try to simplify the task by introducing additional assumptions that are valid in many surveillance scenarios, namely a static camera and a single planar ground plane.

In this paper we propose a robust model for automatically adapting a person detector to an unknown ground plane. The adaptation phase is based on statistics about detection results received from the detector itself. It densely scans a few frames of the sequence at a large number of scales and locations. This information is used for estimating the specific scene scales. In the scene specific detection phase, the search space for this detector is pruned and thus an improvement of computational time and accuracy is achieved.

2 AUTOMATIC SCENE ADAPTATION

By focusing on visual surveillance scenarios where static cameras observe areas containing a single planar ground plane, the general problem of scene adaptation can be simplified. It is assumed that objects of interest are of approximately equal size and that they are located on the ground plane. Therefore the object size depends on the projected position in the image coordinate system. If the camera is mounted horizontally with respect to the ground plane, i.e. there is no camera roll, the size of the object is solely linearly dependent on its vertical position in the image.

Our approach aims at automatically estimating this relationship based on robust statistics about detection results. The goal is to get by with only a small number of detections in a few video frames. Additionally, the usage of a single frame detector avoids a dependency on successful object tracking which becomes difficult in scenes with severe occlusions, rapid appearance changes and crowds. Our proposal to improve a person detector is summarized in Figure 1. First, the detector densely scans sample frames of the input video at multiple locations as well as scales and collects detection results and their confidence scores, if available. Second, based on these detection results the *scene scales are estimated* during the adaptation phase. Third, this information is used for *scene specific person detection* by pruning the search space which in turn provides a computational speedup and higher detection accuracy.

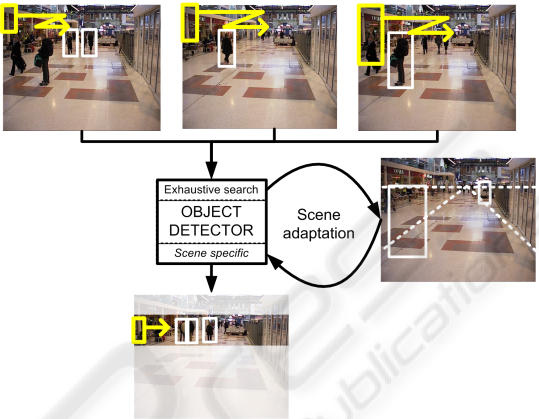


Figure 1: Scene adaptation. Positive detections (white) are collected by exhaustive scanning in multiple scales with a sliding window. After adaptation to the scene, the optimized detector scans the different image areas with detection windows of appropriate scale.

2.1 Scene Scale Estimation

For collecting the person detections we use the publicly available implementation[1] of the histograms of oriented gradients based pedestrian detector from Dalal&Triggs (Dalal and Triggs, 2005). This detector achieves state-of-the-art performance on full-body pedestrian detection (Dollár et al., 2009). For each input image the detector classifies detection windows at multiple scales and locations into 'no pedestrian' or 'pedestrian' each with a confidence score. The input data to the scene scale estimation is a collection of positive person detection results. Specifically, it consists of the persons' feet positions (x and y image coordinate of the bottom center of each detection), the height of the detections and their classification scores. Obviously these observations may contain errors especially in cluttered background and difficult illumination conditions. The task is to robustly estimate the object scale as a linear function of x and y in the presence of false-positive errors. Since noisy data strongly influences linear regression we propose to remove the outliers by fitting a plane into the 3D point cloud using RANSAC (Fischler and Bolles, 1981).

[1] INRIA Object Detection and Localization Toolkit for Windows (http://pascal.inrialpes.fr/soft/olt/), source code from http://www.computing.edu.au/ 12482661/hog.html

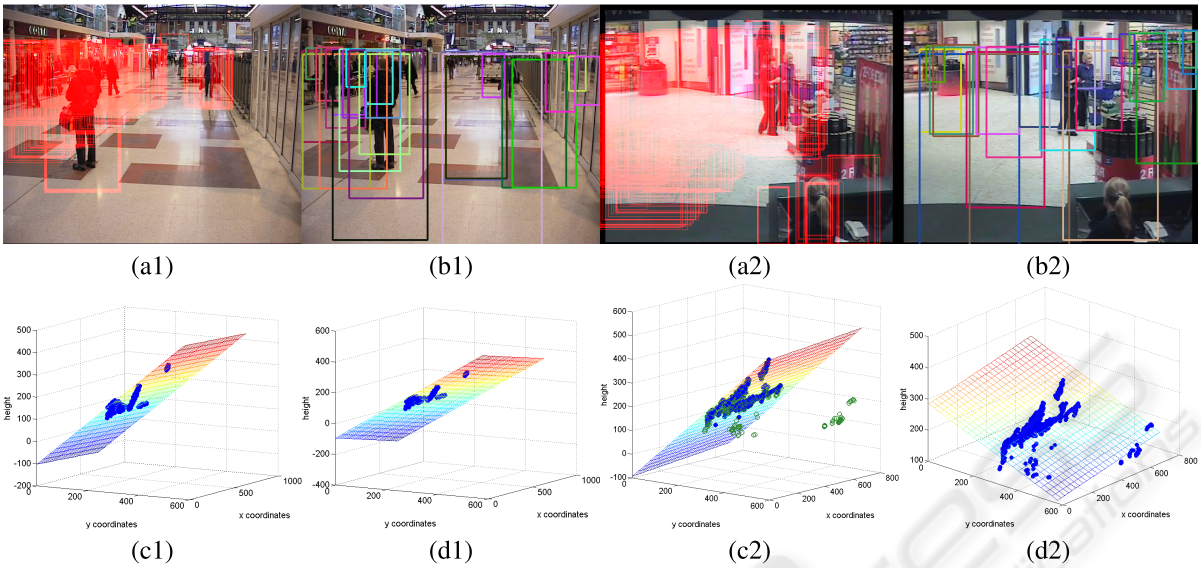


Figure 2: Scene scale estimation for example data with (*2) and without (*1) outliers. Accumulated detections from original detector during adaptation phase (a*), sampled scene specific detection windows obtained from proposed scale estimation (b*), estimated scale model using proposed approach (c*), baseline scale model using robust regression (d*). In the presence of outliers (green circles in c2) a proper scale model is only obtained from the proposed approach. Best viewed in color.

RANSAC is a method for estimating the parameters of a model that optimally fits data with many outliers. The critical threshold value t for determining when a data point fits the model was set to $\frac{1}{4}th$ of the average observed person height which allows for a certain variance in person heights. Subsequently, the linear scale model is robustly fitted on the remaining inliers by taking into account their confidence scores, i.e. considering them as weights. The idea is to downweight the influence of an unreliable observation on the fit. For that purpose we obtain the weighted least-squares solution to the linear system

$$h(x,y) = b(1) * x + b(2) * y + b(3)$$

where - after solving the set of linear equations - b is a vector of size 3 containing the 3D plane coefficients and h represents the estimated scale function depending on the image coordinates x and y. In the above multiple linear regression problem, the weighting is equivalent to multiplying each observation by its confidence score. The greater the weight given to an observation, the more reliable it is. Figure 2 illustrates the idea of using outlier removal and weighted regression for scene scale estimation by means of two examples. For two different scenes 200 collected detections obtained from the detector by exhaustive search during the adaptation phase are shown (a*). The figures in the subplots (c*) show the results of the proposed scene scale estimation by plotting the height over the x and y image coordinates of the observations and the resulting linear scale model. For comparison and baseline, results when using a robust multi-linear regression (robustFit in Matlab) are given in subplot (d*). The figures in the subplots (b*) show examples of estimated detection windows after scene scale estimation. The example on the right (*2) contains false positive observations as can be seen in the bottom right corner of the image with the collected detections (a2) and the data plots. The baseline approach estimated an obviously wrong scale model (d2) because the regression method was strongly influenced by these outliers, plotted with green circles. The proposed approach, however, generates a valid scene scale model (c2).

2.2 Scene Specific Person Detection

To demonstrate the benefit of the scene scale estimation for person detection, we extended the existing implementation of the person detector of Dalal and Triggs (Dalal and Triggs, 2005). This detector densely scans each input image at a large number of possible scales and locations with detection windows of 128x64 pixel size. To this end, the gradients of the image are computed from a scale space pyramid. By default, the pyramid starts at scale 1.0 and gets increased by 5% until the size of the detection window exceeds the dimension of the input image. Subsequently, all detection windows are classified according to their feature descriptor (histogram of oriented

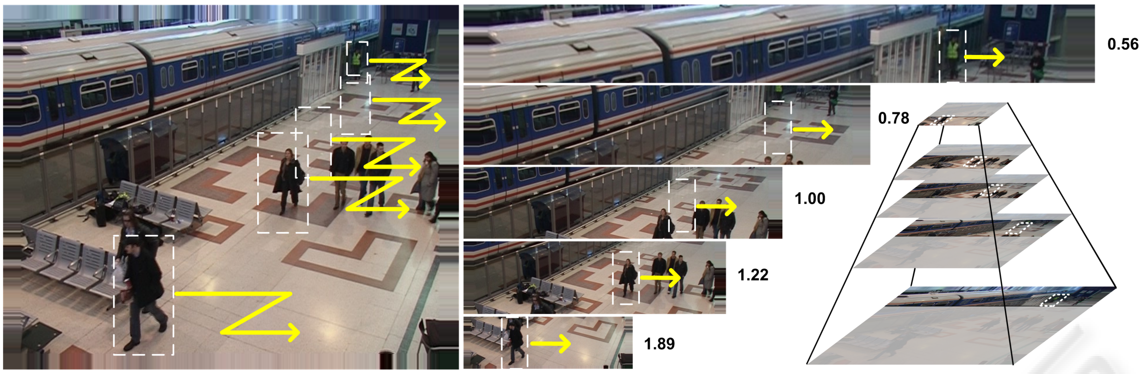


Figure 3: Pruned scale space shown for 5 exemplary scales. Instead of densely scanning the input image (left) at each location and numerous scales, only a subset of the scale space pyramid needs to be processed (center) if scene scale information is used. The bigger part of the scale space remains unprocessed (shaded area on the bottom right).

gradients).

The basic idea of scene specific person detection in videos is to restrict the detection area to the relevant parts in the scale space. A list of detection windows is constructed where for each detection window the scale and location is specified as a result of the scene scale estimation. We extended the existing implementation so that it accepts this list via the newly added command line option (*-sc*). Attention is paid to the fact that the locations and dimensions of the detection windows need to be aligned on a spatial grid because the base implementation tries to cache the feature descriptors for performance reasons. Every different scale involves a preprocessing step where the image is rescaled accordingly, followed by a computation of the image gradients. The benefit of the restricted scale space is that the preprocessing is only made in relevant image parts and scales (see Figure 3).

Summarizing, the scene estimation entails the following performance improvements. First, for each scale the number of detection windows subject to classification is generally reduced. Second, only the relevant scales need to be processed. The number of relevant scales is typically smaller than with exhaustive search. Third, the preprocessing step of each scale speeds up since only subparts of the image are analyzed.

3 EXPERIMENTS AND RESULTS

This section presents evaluation results of the improved scene specific person detector on a variety of different datasets. Figure 5 shows qualitative results on 9 different examples from the i-LIDS(UK Home Office, 2008) and various PETS datasets which are commonly used for benchmarking of detection and tracking systems. For each of the different videos the scene scale is estimated using detection results obtained in the first 150 to 500 video frames (the exact number is shown in the image captions in brackets). Figure 5 demonstrates that false positive detections at wrong scales are generally eliminated if the detector uses the scene scale information (i-LIDS camera 1 and PETS 2009 view 002). The original detector scans the image at scale 1.0 and above, whereas the scene specific detector analyzes the priorly estimated relevant scales. Thus a higher recall is achieved as can be seen from a better detection of people at smaller scales (i-LIDS and PETS 2006).

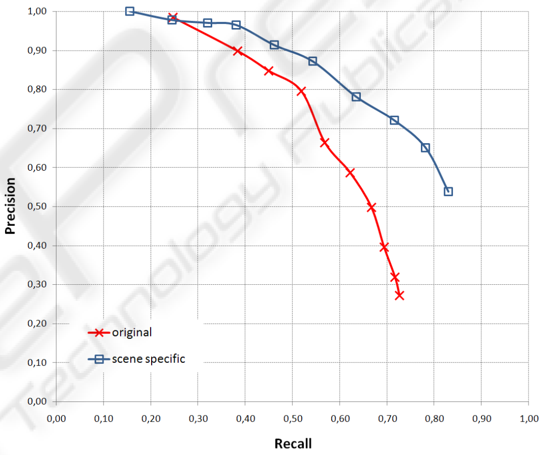
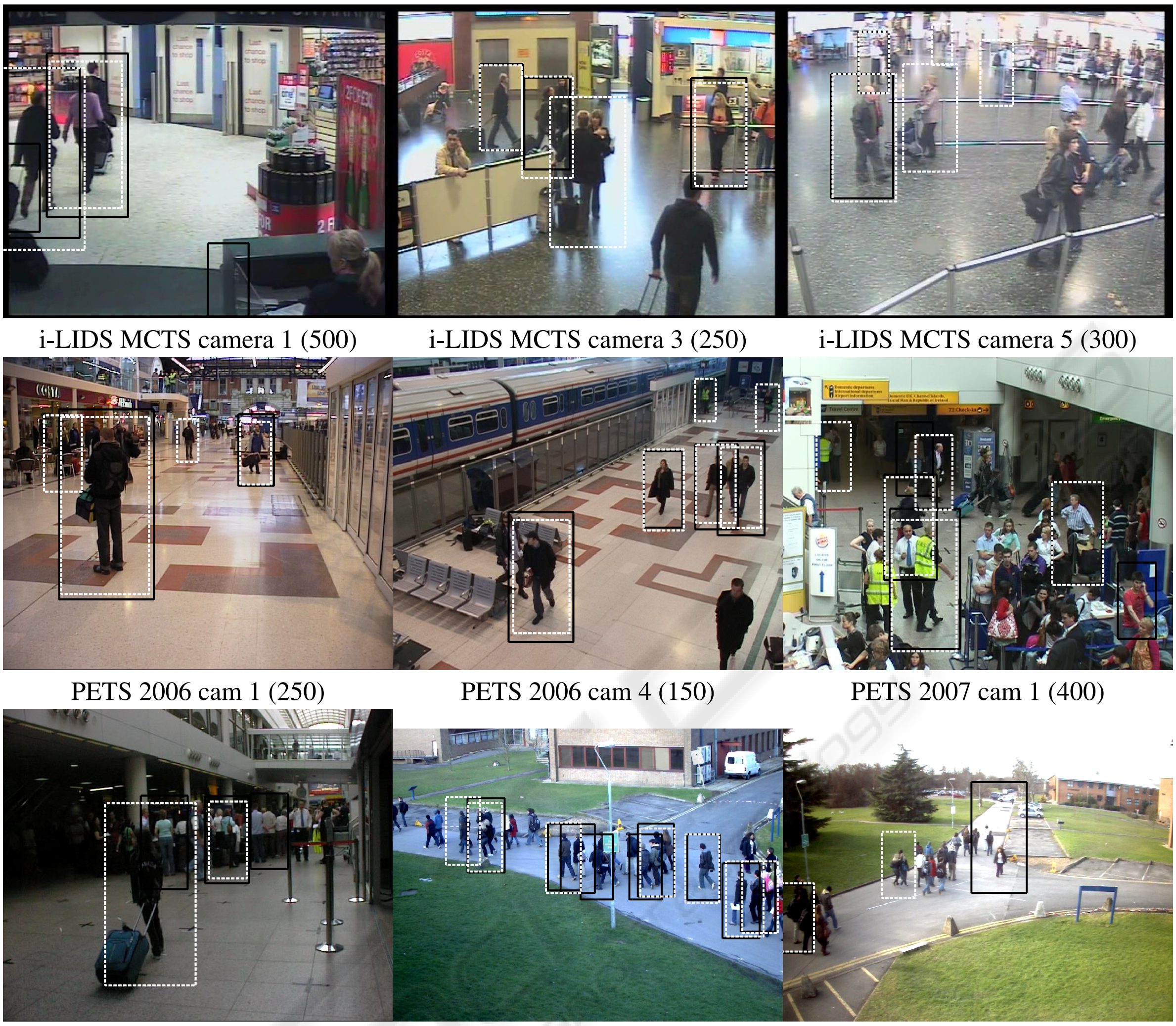


Figure 4: Precision recall graph for varying discrimination thresholds of the scene specific and original person detector.

For qualitative evaluation we compare the results of the scene specific detector and the original detector. To this end, the number of recognizable persons (ground truth), true positive and false positive detections are manually determined for 50 test images of each of the 9 videos. These test images are randomly



i-LIDS MCTS camera 1 (500)

i-LIDS MCTS camera 3 (250)

i-LIDS MCTS camera 5 (300)

PETS 2006 cam 1 (250)

PETS 2006 cam 4 (150)

PETS 2007 cam 1 (400)

PETS 2007 cam 4 (400)

PETS 2009 view 001 (200)

PETS 2009 view 002 (200)

Figure 5: Different test images and detection results obtained from proposed scene specific person detector (white dashed) and the original detector (solid black). The number of images used for the scene scale estimation is given in brackets.

taken from the parts of the videos that have not been used for estimating the scene scale. The criteria for recognizable persons is a person size greater than or equal to 90 pixels and an unoccluded view of at least 80% of the full body. Detections in over-crowded scenes are not taken into account. Further, the discrimination threshold (SVM) of the detectors is varied by using a set of 10 thresholds (0.5, 0, -0.25, -0.5, ..., -1.75, 2.0). In total, detection results for 9000 images (2 detectors * 9 sequences * 50 test images * 10 thresholds) are evaluated. The mean difference in precision and recall is demonstrated in Figure 4. The scene specific detector increases the maximum recall by 10% to 83%. As a result of the increasing number of false positive detections at lower thresholds, the precision of both detectors generally decrease while higher recall values are obtained. Yet, for the scene specific detector the threshold can be lowered with less significant loss in precision. Using the scene scale estimation the improvement in recall is 10% at the precision of 80%, and at the recall of 70% the precision is increased from 38% to 73%.

A comparison of computational performance parameters between the original and the scene specific detector is given in Table 1. It shows the average number of detection windows, scales and run-time performance measured on 100 random images taken from each of the 9 sequences shown in Figure 5.

Table 1: Comparison between the scene specific detector (center column) and the original detector using exhaustive search (left). For achieving comparability, the scene specific detector is also applied on scale 1.0 and above (right).

	exhaust.	scene	scene
scales used	all ≥ 1	relevant	rel. ≥ 1
nr. det. windows	48016	3672	2409
nr. scales	33	29	16
time preproc.	4.82	3.64	1.62
time analysis	10.66	4.26	1.97
time total (sec.)	15.48	7.90	3.59

It has to be noted that the original detectors scans the image at all possible scales greater than 1.0 with an increment of 5%. The proposed scene specific detector analyzes all relevent scales (with the same scale increment), which may include scales smaller than 1.0. To enable direct comparison Table 1 also gives results for the proposed scene specific detector applying a same minimum scale of 1.0. Using the scene scale estimation the number of detection windows and the number of processed scales can be significantly reduced resulting in an average computational speedup by a factor of 4. The increased run-time performance is mainly due to the reduced number of scales and locations at which the feature descriptors have to be computed. Since the base implementation already caches and reuses priorly computed descriptors the 20-fold reduction of the number of detection windows only leads to a 5-fold reduction of analysis time.

4 CONCLUSIONS

A robust approach for automatically adapting a detector to an unknown planar scene is described. Experiments on a variety of datasets demonstrate that scene specific detection gives a speed-up by a factor of 4 and a significant improvement in precision and recall compared to an existing person detector. The Matlab implementation of the scene scale estimation and the code changes to the original person detector in C++ (Dalal and Triggs, 2005) are made available for download [2]. One open issue is the number of observations that are needed for a robust scene scale estimation. Although theoretically only few (3) good detections are required for a planar scene model, the estimate gets more reliable the more detections are available. In our experiments promising results were obtained using a few hundred detections. If many observations are available it is preferable to sample the most probable detections (according to the detector's confidence score) with a large coverage of the image area. Given the low computational complexity of the scene scale estimation an incremental application of the approach is proposed.

ACKNOWLEDGEMENTS

The authors would like to thank their colleague Georg Thallinger, Helmut Neuschmied and Werner Haas. The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 - Challenge 2 - Cognitive Systems, Interaction, Robotics - under grant agreement n216465 - SCOVIS.

REFERENCES

Breitenstein, M. D., Sommerlade, E., Leibe, B., van Gool, L., and Reid, I. (2008). Probabilistic parameter selection for learning scene structure from video. In *BMVC*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2009). Pedestrian detection: A benchmark. In *CVPR*.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*.

Greenhill, D., Renno, J., Orwell, J., and Jones, G. A. (2008). Occlusion analysis: Learning and utilising depth maps in object tracking. *Image Vision Computing*.

Hoiem, D., Efros, A. A., and Hebert, M. (2006). Putting objects in perspective. In *CVPR*.

Renno, J. R., Orwell, J., and Jones, G. A. (2002). Learning surveillance tracking models for the self-calibrated ground plane. In *BMVC*.

Stalder, S., Grabner, H., and van Gool, L. (2009). Exploring context to learn scene specific object detectors. In *Performance Evaluation of Tracking and Surveillance workshop at CVPR*.

UK Home Office (2008). i-LIDS multiple camera tracking scenario definition.

Zhu, L., Zhou, J., Song, J., Yan, Z., and Gu, Q. (2008). A practical algorithm for learning scene information from monocular video. *Optics Express*.

[2] http://scovis.joanneum.at/sceneadaptation