Scene specific people detection by simple human interaction

Scene specific people detection by simple human interaction

Matteo Munaro
Dept. of Information Engineering
University of Padova
matteo.munaro@dei.unipd.it

Angelo Cenedese
Dept. of Management and Engineering
University of Padova
angelo.cenedese@unipd.it

Abstract

This paper proposes a generic procedure for training a scene specific people detector by exploiting simple human interaction. This technique works for any kind of scene imaged by a static camera and allows to considerably increase the performances of an appearance-based people detector. The user is requested to validate the results of a basic detector relying on background subtraction and proportions constraints. From this simple supervision it is possible to select new scene specific examples that can be used for retraining the people detector used in the testing phase. These new examples have the benefit of adapting the classifier to the particular scene imaged by the camera, improving the detection for that particular viewpoint, background, and image resolution. At the same time, positions and scales, where people can be found, are learnt, thus allowing to considerably reduce the number of windows that have to be scanned in the detection phase. Experimental results are presented on three different scenarios, showing an improved detection accuracy and a reduced number of false positives even when the ground plane assumption does not hold.

1. Introduction

Traditional video content analysis research has mainly focused on realizing fully automatic algorithms that do not consider the interaction with a human agent. This choice derives from the fact that the human behavior is variable and not fully predictable and it is sometimes unfeasible to insert a man in the loop of computer vision procedures. On the other hand, though, the capabilities of a state of the art intelligent video surveillance system are far from equating those of a man.

In the last years, many appearance-based people detectors have been proposed [3, 9, 14, 4, 2, 1], whose performance greatly depends on size and variability of the training sets and on the learning procedure that is applied. With this kind of approaches it is difficult to train a classifier that is invariant to changes in viewpoint, illumination, image resolution and background [5]. Furthermore, a sliding window approach [3] is usually adopted for the detection phase, thus meaning that tens of thousands of windows (e.g. for a 720x576 pixels image) are scanned for searching for people at every location and scale, if no additional high level information is exploited. This procedure leads to prohibitive running times and a non negligible number of false positives per image.

In a static camera context some solutions have been proposed for limiting the analysis only to those parts of the image and scales where a person is possible to be present. Background subtraction [14] and probabilistic techniques [11, 8] have been widely applied for extracting the foreground, but with well known limitations, such as merging of close targets or of targets with their shadows and the need of good contrast between foreground and background.

Geometric information about the scene, such as the presence of a ground plane, has also been exploited [7] in order to limit the size of people in the image. However, this assumption only holds if a single ground plane exists, but this is not the case of scenes featuring, for example, changing slopes or stairs.

In this work, we propose a general technique that, at the cost of a simple interaction with a human agent, can both improve the accuracy and reduce the computational time of a state of the art people detector without any geometric assumption on the imaged scene.

1.1. Related Work

Most of the research done on scene specific detectors deals with the problem of selecting new training examples for retraining the detector in order to make it adapt to the specific conditions of viewpoint, background and resolution. These new examples must be reliable and introduce complementary information with respect to that already present in the training set. Along this line, for example, Nair and Clark [10] exploited a background subtraction algorithm to select some person examples from a training video. However, due to the inaccuracy of background subtraction, this procedure introduced also bad examples in the

training set, thus leading the people detector to easily drift. Wang and Wang [13] combined different cues for the automatic selection of new examples obtaining good accuracy improvement for a particular traffic scenario and assuming an eagle-eye perspective for the camera. Stalder *et al*. [12] exploited a tracking procedure for improving detection in presence of occlusion, but they also assumed the presence of a ground plane. In Gualdi *et al*. [6] a human agent is requested to select the good detections among the labels provided by an automatic people detector, so as to increase the reliability of a subset of validated examples that can be used to retrain the detector. However, since the same detector is used for both the training and test phases, the selected people can result as not enough informative examples for the retraining phase. Furthermore these examples are employed only in the last stages of a cascade classifier, thus resulting in a reduced number of false positives, but not of false negatives. In addition, the algorithm in [6] also relies on the ground plane assumption to prune the number of search windows.

1.2. Our Approach

In the approach described in this paper we select new training examples by running a basic people detector based on background subtraction on an appropriate training video where people appears roughly in all the locations they can access in the scene and asking a human agent to validate the presented results. Therefore, unlike Gualdi *et al*. [6], we use a different detector for the training and the test phases and we exploit the gathered examples to reduce both the false positives and the false negatives rates. As an additional result of the training phase, positions and scales where people can be found in the image are learnt and used in the detection phase, thus allowing a considerable reduction of computational time and false detections. A significant advantage respect to previous work is that the proposed technique does not make any geometric assumption on the scene, thus it is applicable to every type of scene, even where no planar ground plane is present.
The paper is organized as follows: in Section 2 the human interaction procedure and the useful information that can be learnt are described. In Section 3 experiments on three different scenarios are reported, showing improvements both in detection accuracy and computational time. Conclusions are presented in Section 4.

2. Human interaction and its benefits

2.1. Validation procedure

In order to improve the performance of an appearance-based people detection system, we request a human agent to validate some detections coming from a basic people detector that can easily run in real time. The basic detector we use here performs background subtraction and applies a threshold and some constraints on the dimension and proportions of a blob. As we said in Section 1, these kind of techniques suffer many problems, that is the resulting detections can contain also some shadows, or only part of a person because the foreground has not enough contrast with the background, or more persons can be merged into the same blob if they are too close. Thus the blobs that pass this preliminary test are provided to a human agent for validation. The operator's task consists in choosing the detections that contain only one whole person by simply clicking inside the corresponding windows, which is the type of examples fed to the algorithm that train the people classifier. The human agent can also properly set the threshold of the background subtraction algorithm in order to have the best separation between background and foreground. Both the proposed tasks are fast to be performed, can be executed by non expert people, and do not require particular precision in the selection. In Figure 1(a) the result of background subtraction after thresholding and some morphological operations[1] is reported for a video frame of the *PETS 2006* dataset[2]. After the foreground pixels are separeted in connected components, a simple constraint on blob proportions is applied in order to select standing people candidates[3]. Figure 1(b) shows in red (solid line) the bounding boxes of the blobs that passed this preliminary test. Once a blob is validated by the human agent, its bounding box is enlarged (in green (dotted line) in the figure) in order to contain also a proper contour around the person, similarly to the INRIA[4] examples used by Dalal and Triggs [3]. Then that portion of the image is resized to the canonical dimensions of 128x64 pixels and saved as a new *scene-specific* positive example.

(a) Foreground mask. (b) Proposed (red/solid) and validated (green/dotted) detections.

Figure 1. The acquisition steps for the *scene-specific* examples.

2.2. Scene-specific data

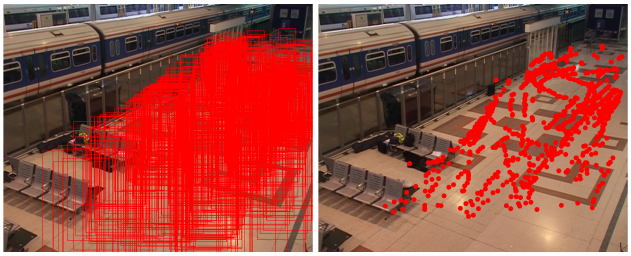
The *scene-specific* examples are used, together with the default ones, to retrain the people detector in order to in-

[1] An opening operation can be enough to remove salt and pepper noise.
[2] This dataset can be freely downloaded at http://www.cvg.rdg.ac.uk/PETS2006/data.html.
[3] E.g. blobs with a width-height ratio between 0.2 and 0.7 and not connected to the image borders.
[4] http://pascal.inrialpes.fr/data/human.

crease its accuracy in detecting people from that particular scene. These new examples are uncorrelated from the default ones, since they are not selected by the people detector we use in the test phase, so they are more likely to be informative for retraining. We randomly select also some *scene-specific* negative examples from the background image and add them to the default ones.

Further useful information can be jointly derived by the validation procedure described in Section 2.1. In fact, every validated window provides also information about the image position and scale at which a person can appear. As a result, if the human agent validates detection windows near to every possible location of people in the image, the detection algorithm can learn where to search for people and at which scale.

For instance, in Figure 2(a) all the detection windows validated by a human agent from 480 frames of the *PETS 2006* video are drawn, while in Figure 2(b) the correspondent windows centroids are reported. As it can be seen, almost all the area where people can be found is covered by the selected detection windows, thus they give a good indication of where to search for people in the image.



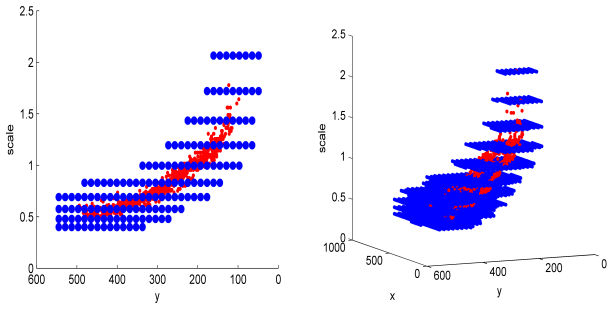
(a) Selected windows. (b) Centroids of the selected windows.

Figure 2. All selected positive examples for the *PETS 2006* video.

2.3. Detection phase

We use the sliding window technique for densely analyzing an image at different scales where every scale differs from the previous one by a constant scale factor. In a (x,y,$scale$) representation of the search space, where the axis are the x and y position of the windows centroid in the image and the window scale, the detection windows to be analyzed at every scale are points that lie on planes parallel to the xy plane. In Figure 3 the windows corresponding to the *scene-specific* examples of Figure 2(a) are represented as red points in this 3D space. A scale value of 1 corresponds to a detection window of standard dimension, 128x64 pixels. Values that are larger than 1 represent smaller windows, while values smaller than 1 refer to bigger windows. This figure well highlights that people's size in the image increases with the y coordinate.

As a pruning strategy for the windows to be classified in the detection phase we propose to analyze, at every scale, only those windows that lie in a 3D neighborhood of those selected by the human agent. This neighborhood is defined by setting a proper threshold on the Euclidean distance after having normalized the three axis. So far this threshold has been empirically set. In Figure 3 the windows that are analyzed are represented as blue points around the red ones. These portions of plane can be depicted through a 3D binary



(a) Lateral view (b) 3D view

Figure 3. Windows selected in the validation phase (red points) and windows to be analyzed in the detection phase (blue points).

matrix that is set to 1 in correspondence to the centroids position of windows that could contain a person at a particular scale. An example of this matrix is illustrated, scale by scale, in Figure 4 for the *PETS 2006* video when choosing 1.2 as scale stride. This matrix can be easily computed offline and used in the detection phase for analyzing only those windows corresponding to matrix points (x,y,$scale$) set to 1.

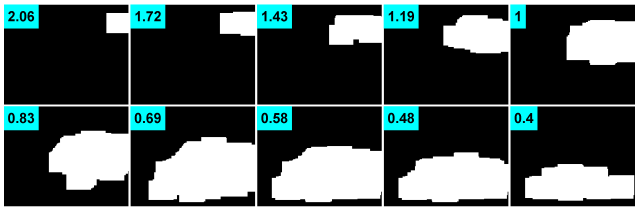


Figure 4. *Scale matrix* illustration: the white zone represents where the centroids of the windows to be analyzed lie; the numbers represent the scale values.

This pruning procedure does not rely on any geometric assumption (e.g. the existence of a ground plane), so it is fully generic and applicable to every type of scene.

3. Experimental results

We show the performance of the described approach with three videos corresponding to different scenarios imaged by a static camera. Every video sequence has been partitioned into one section used for selecting the *scene-specific* examples through foreground extraction and hu-

man validation, and the remainder used for testing. As for the people detector, we used Dollár's implementation of HOG[5] and the same procedure and parameters described by Dalal and Triggs [3] for training the default detector. However, since the proposed technique is generic, any other appearance-based pedestrian detectors can be used.
Three different approaches have been compared with the test videos:

- *generic*: people are densely searched in all the image at every scale with the HOG detector trained with the default examples;
- *semi scene-specific*: the pruning procedure is used for limiting the number of windows to be analyzed, but the default detector is used for classification;
- *scene-specific*: both the pruning procedure and the re-trained detector with the *scene-specific* examples are used for detection.

3.1. *PETS 2006* video

The first results presented here refer to a video of the *PETS 2006* dataset. In this scenario the camera is tilted of about 30 degrees with respect to the horizontal, so people are seen from above, while the most part of the default training examples that are fed to our learning algorithm are depicted from a frontal view. In addition to these default examples 800 *scene-specific* examples have been selected from a part of the video not used for testing. Some of these new examples are reported in Figure 5.

Figure 5. Some of the *scene-specific* examples acquired for the *PETS 2006* video.

We also added 500 negative windows to the negative training set, extracted by the background image. This image has been obtained by means of a median filter applied pixel-wise to the video frames. It has also been added to the set of images used for the bootstrapping procedure [3].
The test set is composed of 1000 frames that had not be used for the validation phase. Figure 6 reports a comparison between the *generic* detector and the *scene-specific* when setting the confidence threshold to -0.5, the scale stride to 1.05 and the default parameters for the HOG descriptor [3]. The improvement obtained with the *scene-specific* approach is remarkable in terms of new persons detected, increased confidence in the detection, and false positives avoided. The first two features are mainly due to the adaptation of the classifier to the current point of view, while the last is strongly related to the restriction of the people search space. For what concerns the computational time, the *scene-specific* detection allows to save 75% of the computational time. The performance of the *generic*, *semi scene-specific* and *scene-specific* detectors can be better understood with the DET curves [5] obtained by varying the confidence threshold and reported in Figure 7. We adopted the PASCAL criterion [5] for comparing the detection results with the ground truth and in the two axis the number of *False Positives Per Frame* and the *False Rejection Rate* are shown. At 10^{-1} FPPF 10% more people are detected by the *scene-specific* detector while formerly neglected by the *generic* one.

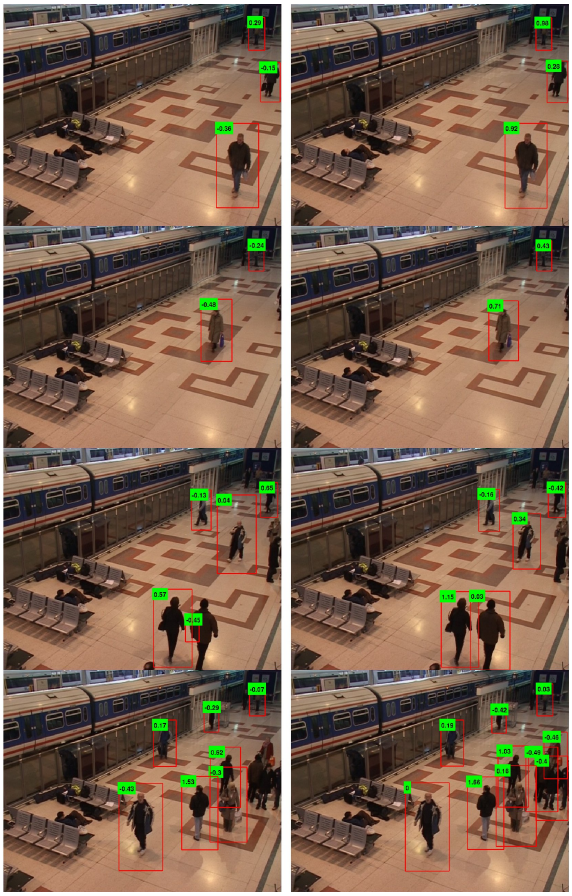


Figure 6. Detection results on the *PETS 2006* video with the *generic* (left) and *scene-specific* (right) detector. The confidence threshold is set to -0.5.

3.2. *Prato* video

The second video we used for testing refers to a scene with a changing slope, where the ground plane assumption

[5]Contained in his Matlab toolbox http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html.

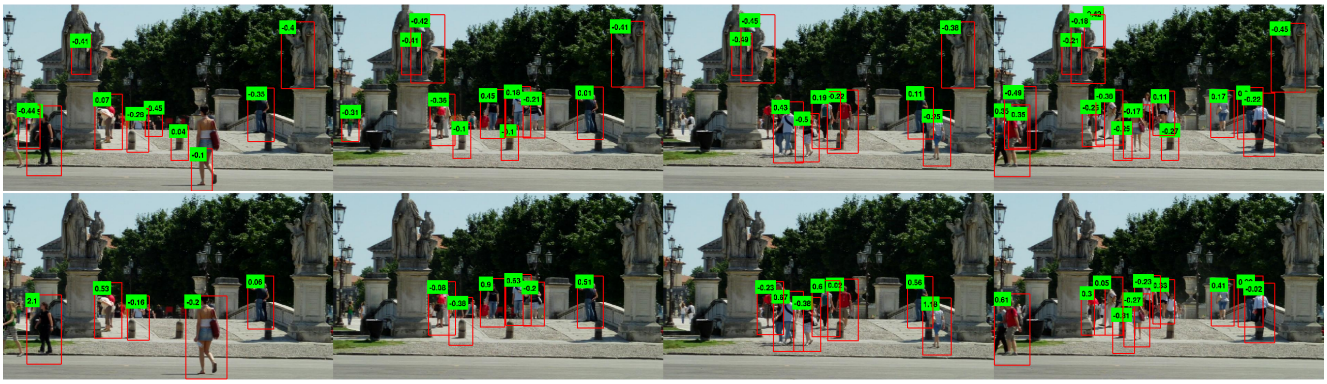


Figure 10. Detection results on the *Prato* video with the *generic* (first row) and *scene-specific* (second row) detector. The confidence threshold is set to -0.5.

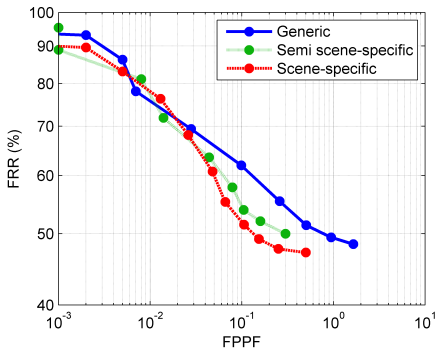
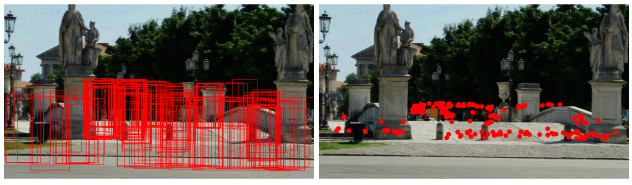


Figure 7. DET curves for the *PETS 2006* video.

does not hold. From a part of this video we acquired 416 new positive examples and 500 negative examples. Figure 8 illustrates the position of all the selected positive examples in the image and Figure 9 shows the content of some of these windows.

The test set here is composed by 1000 frames at a resolution of 960x540 pixels. In Figure 10 the detection results of the *generic* and *scene-specific* detectors are compared and the DET curves are reported in Figure 11. From the image comparison it is easy to check the effectiveness of the pruning strategy in removing some false positives on the statues or that have dimensions not compatible to those of a person. The increased confidence on the detection windows



(a) Selected windows. (b) Centroids of the selected windows.

Figure 8. All selected positive examples for the *Prato* video.

Figure 9. Some of the *scene-specific* examples acquired for the *Prato* video.

containing a person can also be noticed. Finally, the pruning strategy allows here to save the 78% of the computational time.

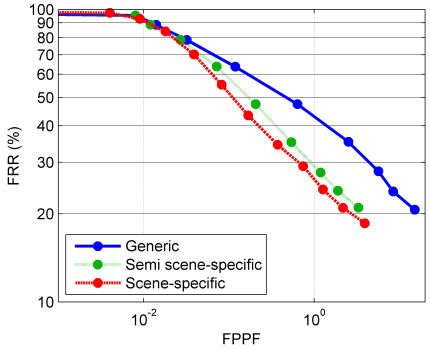
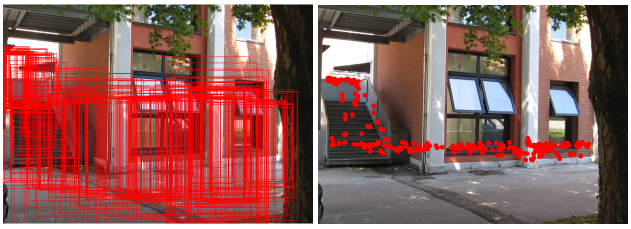


Figure 11. DET curves for the *Prato* video.

3.3. *Stairs* video

In the *Stairs* video, people moving around some emergency stairs are present in the scene. This is a another typical video surveillance scenario where the ground plane assumption does not hold, while our procedure is applicable. From a training video we acquired 600 new positive examples and 500 negative examples. Fig. 12 illustrates the position of all the selected positive examples in the image, while the DET curves for this video are reported in Fig. 13.

Even for this video a clear performance improvement is obtained with the *scene-specific* approach and 60% of the



(a) Selected windows (b) Centroids of the selected windows

Figure 12. All selected positive examples for the *Stairs* video.

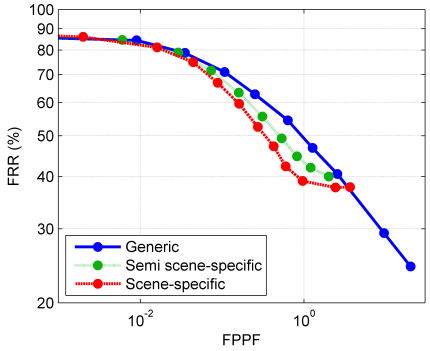


Figure 13. DET curves for the *Stairs* video.

	PETS 2006	***Prato***	***Stairs***
Gain factor	4x	4.4x	2.4x

Table 1. Gain factor obtained with the *scene-specific* detection respect to the *generic* one in terms of computational time.

computational time can be saved. A summary of the computational gain for every video is reported in Table 1.

4. Conclusions and future work

This paper proposed a generic procedure for training a *scene specific* people detector by exploiting simple human interaction. This technique works for any kind of scene viewed by a static camera and allows to considerably increase the performances of an appearance-based people detector and reduce the computational time. It exploits human validation for ensuring good quality of the *scene-specific* examples added for retraining and background subtraction as a preprocessing step to make the human validation a very simple and fast task. This technique has been proved to work in generic scenarios, even when the ground plane assumption does not hold.

As a future work a tracking algorithm could be introduced to improve the accuracy of background subtraction. Moreover a new rule could be studied in order to automatically set the threshold used for selecting the valid neighborhood around the validated examples.

Acknowledgements: the authors would like to thank Prof. P. Perona for the stimulating and fruitful discussions on the subject and Videotec S.p.a. for the support throughout this work. This activity contributes to the seed project R3D of the Department of Information Engineering - University of Padova.

References

[1] Y. Abramson, Y. Freund, and R. Pelossof. Semi-automatic visual learning (seville): a tutorial on active learning for visual object recognition. In *CVPR 2005 Tutorial*. 1

[2] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *Proceedings of the 11th European conference on Computer vision: Part VI*, ECCV'10, pages 168–181, Berlin, Heidelberg, 2010. 1

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005.*, volume 1, pages 886–893, june 2005. 1, 2, 4

[4] P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *BMVC*, 2010. 1

[5] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition*, pages 304–311, 2009. 1, 4

[6] G. Gualdi, A. Prati, and R. Cucchiara. Contextual information and covariance descriptors for people surveillance: an application for safety of construction workers. *J. Image Video Process.*, 2011:9:1–9:16, January 2011. 2

[7] D. Hoiem, A. A. Efros, and M. Hebert. Putting Objects in Perspective. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2006. 1

[8] V. Kruger, J. Anderson, and T. Prehn. Probabilistic model-based background subtraction. In *Image Analysis and Processing - ICIAP 2005*, volume 3617 of *Lecture Notes in Computer Science*, pages 180–187. 2005. 1

[9] Z. Lin, L. S. Davis, D. S. Doermann, and D. DeMenthon. Hierarchical part-template matching for human detection and segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV'07)*, pages 1–8. IEEE, 2007. 1

[10] V. Nair and J. Clark. An unsupervised, online learning framework for moving object detection. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2004. 1

[11] J. Rittscher, J. Kato, S. Joga, and A. Blake. A probabilistic background model for tracking. In *European Conference on Computer Vision*, pages 336–350, 2000. 1

[12] S. Stalder, H.Grabner, and L. V. Gool. Exploring context to learn scene specific object detectors. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Miami, USA*, pages 63–70, June 2009. 2

[13] M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *Computer Vision and Pattern Recognition, 2011. CVPR 2011.* 2

[14] T. Zhao, R. Nevatia, and B. Wu. Segmentation and tracking of multiple humans in crowded environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1198–1211, 2008. 1