People Detection using Color and Depth Images*

Joaquín Salas[1] and Carlo Tomasi[2]

[1] Instituto Politécnico Nacional, `jsalasr@ipn.mx`
[2] Duke University, `tomasi@cs.duke.edu`

Abstract. We present a strategy that combines color and depth images to detect people in indoor environments. Similarity of image appearance and closeness in 3D position over time yield weights on the edges of a directed graph that we partition greedily into *tracklets*, sequences of chronologically ordered observations with high edge weights. Each tracklet is assigned the highest score that a Histograms-of-Oriented Gradients (HOG) person detector yields for observations in the tracklet. High-score tracklets are deemed to correspond to people. Our experiments show a significant improvement in both precision and recall when compared to the HOG detector alone.

1 Introduction

The detection of human beings from visual observations is a very active research area. The recent introduction of inexpensive depth sensors that work at frame rate offers new opportunities to address this difficult problem. In this paper, we combine depth and color data from a single sensor to track and classify people. More specifically, we introduce a directed graph whose edges connect chronologically ordered observations. Weights on the graph capture similarity of appearance and closeness in space, and a greedy traversal of the graph produces *tracklets*, that is, chronological sequences of observations that are likely to correspond to the same person. Each tracklet then receives a score from a color-based person detector from the literature [1]. Tracklets with scores exceeding a predefined threshold are deemed to correspond to people. Our experiments show that our strategy reduces the number of detected false positives by a factor of fifty, while increasing the detection of true positives threefold. The rest of the paper is structured as follow. After a brief review of related work, Section 3 describes a method to extract foreground objects using depth information. Then, Section 4 discusses the creation of tracklets, and Section 5 presents results on two color/depth video sequences. Comparison with ground truth data illustrates the

* Thanks to Julian (Mac) Mason for his gentle introduction to the Kinect, including his help for calibrating the sensor and obtaining the first set of images. This work was supported by the Consejo Nacional de Ciencia y Tecnología under Grant No. 25288, the Fulbright Scholarship Board, and the Instituto Politécnico Nacional under Grant No. 20110705 for Joaquín Salas, and the National Science Foundation under Grant No. IIS-1017017 and by the Army Research Office under Grant No. W911NF-10-1-0387 for Carlo Tomasi.

benefits of our approach when compared to HOG detection alone. A concluding Section suggests directions for future research.

2 Previous Work

An account of early efforts on people tracking can be found in [7]. These include the analysis of parts of the body, both internal or external, as well as dynamical characteristics, such as the gait. Some of the first results can be traced back to the Seventies [12], when psychophysical studies [11] showed that humans could perceive people based on pure motion data. Prompted in part by security considerations [20], new techniques, protocols and standards have emerged in the past decade. Some approaches have used silhouettes [5] or body-part matching [14, 21, 23, 22]. The combination of cascades of increasingly complex classifiers has produced fast and robust recognition algorithms [28] for relatively stylized person poses. Features for tracking people include the Scale Invariant Feature Transform (SIFT) [13], [15], Haar-like wavelets [29], shape [33], and Histograms of Oriented Gradients (HOG) [1]. The latter have proven to be particularly successful. To build a HOG descriptor, the window of interest in an image is subdivided into a grid of cells, and a histogram of the orientations of luminance gradients is computed in each cell. The histograms are normalized and concatenated into a single vector for the whole window. A linear Support Vector Machine (SVM) [27] classifies the resulting vectors into person or non-person. This work was later extended [2] to include the use of motion. Motion information had been used in other work as well [29], [6]. SVMs have been used with other descriptors for whole bodies [16] or body parts [19]. Schwartz *et al.* [25] further incorporated texture information.

Some researchers have combined spatial and light intensity information to detect people. For instance, Zhao and Thorpe[34] use a stereo system to segment the silhouettes that are fed to a neural network that detects pedestrians. Xu and Fujimora [32] also extract body silhouettes but with a time-of-flight device. The use of body, the whole or just parts of it, has proven to increase the robustness of the detection and tracking methods. Consider for example the strategy proposed by Muñoz *et al.*[17] where there is the combined use of a face detector and depth information to track people. Javed *et al.* [10] instead combine color with position information inferred from the locations of multiple cameras. In our work, we use similar principles for combining color and position information. However, we work in the field of view of a single color/depth sensor, and derive position information from a depth map through background subtraction. In addition, we also run a HOG classifier on every color frame, and propagate the best scores it generates to all observations in the same tracklet. Thus, we classify one tracklet at a time, rather than one window at a time. While this approach propagates both true positives and false positives, our reliance on the *best* detection result in each tracklet ensures that the HOG classifier is given the opportunity to operate on body poses that fit the HOG model particularly well. The good results of our experiments in Section 5 show the validity of our approach.

3 Detection of Foreground Objects

We first classify the measurements $X = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ from a depth sensor, where $\mathbf{x}_k = [x_k, y_k, z_k]^T$, into background $\mathcal{B}$ and foreground $\mathcal{F}$. To this end, a Gaussian background model is used to detect the foreground by Maximum A Posteriori (MAP) estimation. The resulting foreground points are then grouped into separate objects by connected component analysis. For our purposes, we divide the tridimensional space into equally spaced bins centered at $\mathcal{X} = \{\overline{x}_1, \dots, \overline{x}_a\}$, $\mathcal{Y} = \{\overline{y}_1, \dots, \overline{y}_b\}$, and $\mathcal{Z} = \{\overline{z}_1, \dots, \overline{z}_c\}$ with grid spacing Δx, Δy, and Δz. At the workspace boundaries, the bins extend to either ∞ or $-\infty$. In the following, $\mathcal{N}$ is a function that counts the number of observations that fall into each of the bins of a histogram.

3.1 Planar Background Elimination

Similarly to Vrubel *et al.* [30], we assume that the workspace has either a flat floor or a flat ceiling. Furthermore, we assume that the number of points describing either one of these structures is a significant fraction of the points in the depth map. We then compute the sensor roll and pitch angles that produce a maximum bin value over the marginals on the vertical axis. Specifically, let $h(j, \alpha, \beta) = \mathcal{N}(|\overline{y}_j - y| \leq \Delta y/2)$ be the marginal histogram along the vertical direction after a rotation of the reference system by roll and pitch angles α and β. The rotation that maximizes the number of points in the most populated bin, that is, $(\alpha, \beta) = \arg\max_{\alpha,\beta} \max_j h(j, \alpha, \beta)$, can be estimated using the Nelder-Mead or Simplex method [18]. For efficiency, the points below the floor and above the ceiling are deleted after this rotation.

3.2 Background Model

The Occupancy Grid framework [4] provides a suitable platform for background subtraction. Let $s(\mathbf{x})$ be a foreground/background map for the spatial coordinates $\mathbf{x} \in X$, with $p(s(\mathbf{x}) = \mathcal{F}) + p(s(\mathbf{x}) = \mathcal{B}) = 1$. The probability that a particular space position $\mathbf{x} = [x, y, z]^T$ is part of the background is

$$p(s(\mathbf{x}) = \mathcal{B}|z) \propto p(z|s(\mathbf{x}) = \mathcal{B})p(s(\mathbf{x}) = \mathcal{B}). \tag{1}$$

Similarly to Gordon *et al.* [9], who presented a method to combine dense stereo measurements with color images, we model the background with a mixture of Gaussians and detect the foreground as those points that are more than 3σ away from the nearest background mode.

3.3 Foreground Objects

We extract foreground objects by connected components with 26-connectivity in 3D space, while reasoning about the positions of the detected objects relative to the sensor. Let $\mathcal{H}$ be a histogram constructed out of the points in X, such that

$\mathcal{H}(i,j,k) = \mathcal{N}(|\overline{x}_i - x| \leq \Delta x/2, |\overline{y}_j - y| \leq \Delta y/2, |\overline{z}_k - z| \leq \Delta z/2)$. Let $v(i,j,k)$ be an indicator variable that is 1 whenever $\mathcal{H}(i,j,k) > 0$ and 0 otherwise. Objects correspond to connected components in $v(i,j,k)$. Finally, we eliminate clusters that are smaller than a depth-dependent threshold of the form $\tau(d) = \rho e^{-\nu d}$ that models the fact that the size of an object decreases with its distance d from the sensor. The values of ρ and ν are found by data fitting on training samples. Each output *blob* is given in the form of the tightest axis-aligned box around each component.

4 Combining Detections

To combine measurements of depth and appearance, we use depth for tracking blobs across frames and connecting them into *tracklets*, and we use the HOG detector [1] in one of its available implementations [3, 31] to assign scores to individual blobs. The highest score on each tracklet is then propagated to all the blobs on that tracklet. Blobs that are in tracklets with a score that exceeds a given threshold are classified as people. In this Section we describe our construction of tracklets.

Adapting the framework proposed by Javed *et al.* [10], in our case for a single camera, let k_i^j be a binary indicator for the hypothesis that two observations $O_i = \{\mathbf{f}_i, \mathbf{x}_i, t_i\}$ and $O_j = \{\mathbf{f}_j, \mathbf{x}_j, t_j\}$ belong to the same object. In each observation, $\mathbf{f}$ is the blob color signature[24], $\mathbf{x}$ is the position of the centroid of the points in a blob, and t is the timestamp of the observation. The conditional probability distribution of k_i^j given two observations O_i, O_j is

$$p(k_i^j | O_i, Oj) \propto p(\mathbf{f}_i, \mathbf{f}_j | k_i^j) p(\{\mathbf{x}_i, t_i\}, \{\mathbf{x}_j, t_j\} | k_i^j) p(k_i^j), \tag{2}$$

assuming independence of $\mathbf{f}$ from $(\mathbf{x}, t)$. Lacking further information, we may assume that $p(k_i^j)$ is uniformly distributed. We define

$$p(\mathbf{f}_i, \mathbf{f}_j | k_i^j) \propto e^{-\alpha d(\mathbf{f}_i, \mathbf{f}_j)}, \tag{3}$$

where $d(\mathbf{f}_i, \mathbf{f}_j)$ is the Earth Movers Distance (EMD) [24]. We also define

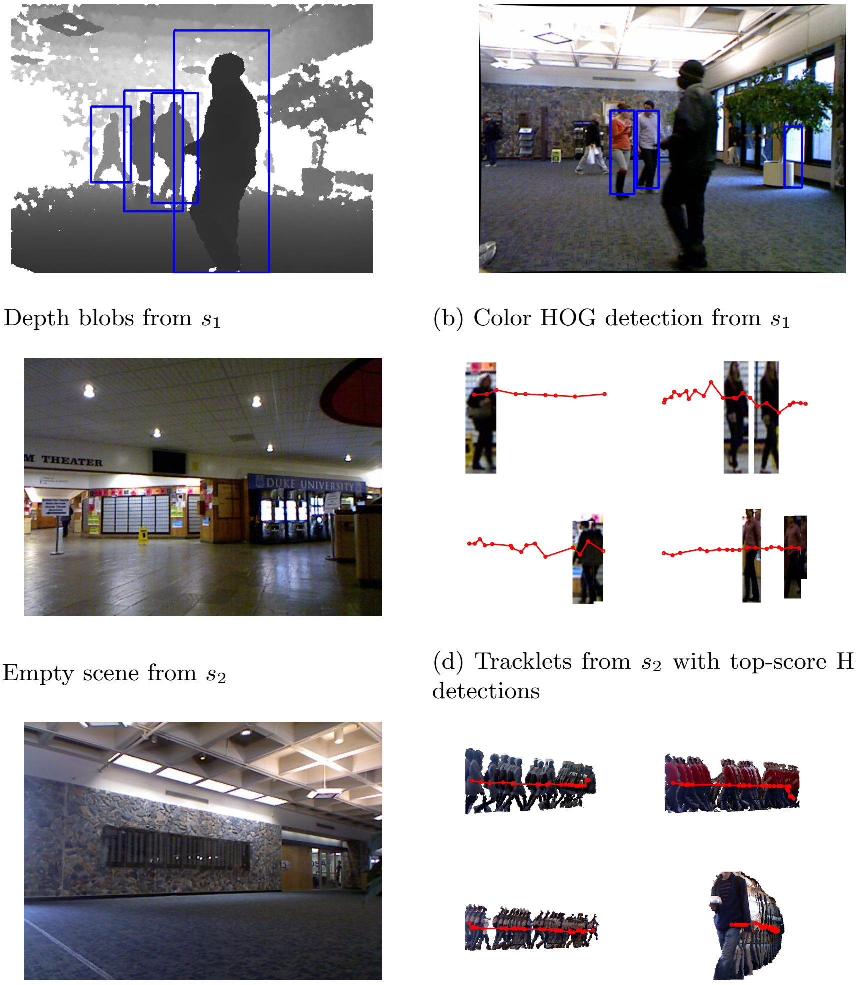
$$p(\{\mathbf{x}_i, t_i\}, \{\mathbf{x}_j, t_j\} | k_i^j) \propto e^{-\beta \|\mathbf{x}_i - \mathbf{x}_j\| - \gamma \|t_j - t_i - \Delta t\|} \tag{4}$$

where Δt is the inter-frame time. We estimate the constants α, β and γ in these expressions through data fitting to training samples.

To compute tracklets, we build a directed graph $G = (V, E, P)$ whose node set V is the set of observations O_i, edge (i,j) in E connects observations O_i and O_j such that $t_i < t_j$, and the weights in P are the probabilities $\pi_{ij} = p(k_i^j = 1 \,|\, O_i, O_j)$, evaluated as explained above. Edges with zero weight are omitted from E. In G, we define *tracklets* as strongly connected paths, constructed greedily as follows: Let i_0 be the oldest observation in V. For each i, let

$$j(i) = \arg \max_{j \in V, (i,j) \in E} \pi_{ij}. \tag{5}$$

A tracklet is the resulting path $i_0\,,\ i_1 = j(i_0)\,,\ i_2 = j(i_1)\,, \ldots$. The path ends when $j(i_n)$ is undefined. We then remove the elements of the tracklet from the graph, and repeat.



(a) Depth blobs from s_1

(b) Color HOG detection from s_1

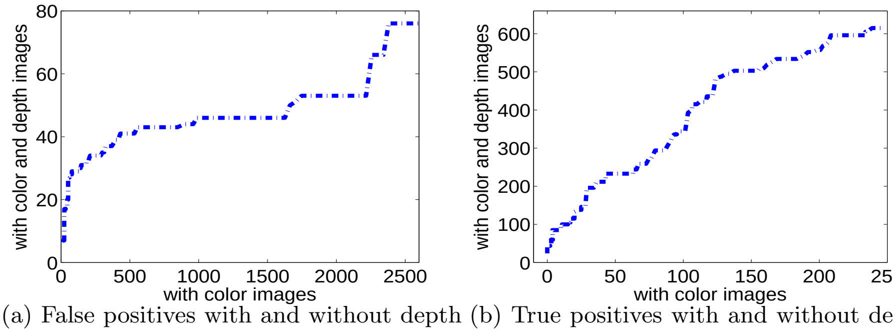
(c) Empty scene from s_2

(d) Tracklets from s_2 with top-score HOG detections

(e) Empty scene from s_3

(f) Tracklets from s_3 with foreground blobs

Fig. 1. People Detection using Color and Depth Images. (a) Bounding boxes of foreground connected components found in a depth frame from s_1. (b) HOG detections in a color frame from s_1. In this particular frame, the HOG finds two people, misses two, and produces a false positive to the right of the plant. (c) A frame from s_2, with no people. (d) Tracklets (red polygonal lines) *from* s_2 with superimposed top-scoring HOG detection results. (e) A frame from s_3, with no people. (f) Tracklets (red polygonal lines) from s_3 with superimposed foreground blobs.



(a) False positives with and without depth information

(b) True positives with and without depth information

Fig. 2. Performance Evaluation. In these graphs, we vary the acceptance threshold τ for the HOG detector from 0 (all pass) to 3.15 (most strict). In each diagram, the horizontal axis is the number of HOG detections on the color images, and the vertical axis is the number of detections with our method. (a) Number of false positive detections (`fp`). For $\tau = 0$, HOG alone obtains 2,594 `fp` and our method yields 76. For $\tau = 3.15$, `fp` is 2 for HOG alone and 7 with our approach. (b) Number of true positive detections (`tp`). For $\tau = 0$, HOG alone finds 245 `tp` and our method finds 615. When $\tau = 3.15$, `tp` is 0 for HOG alone and 16 for our approach. A standard ROC curve [26] would be essentially meaningless, because the multiscale image scan examines 34,981 windows per image, vastly more than the expected number of targets.

5 Experimental Results

Our experiments evaluate the impact of depth information, used as described earlier, on the performance of the HOG person detector. To this end, we captured and processed three sequences s_1, s_2 and s_3 with a Microsoft Kinect sensor [8]. Each sequence contains roughly 2,000 color/depth image pairs at VGA resolution (640×480). Sample frames are shown in Fig. 1. We divided the workspace in cells with grid step $\Delta x = \Delta y = \Delta z = 0.1$m. Using the MATLAB implementation of the Nelder-Mead optimization algorithm [18], we estimated the pitch (β) and roll (α) angles. We used floor points in s_1 and s_3, and ceiling points in s_2. The estimated angles for pitch and roll are -3.4 and -1.1 degrees for s_1, 0.9 and 4.1 for s_2, and -0.9 and 3.4 for s_3 . Only points between 0.1m and 2.5m above floor level are considered for processing.

To construct a model of the background, we chose 20 frames from s_1, 80 from s_2, and 160 from s_3, consecutive and without people. To detect people, we used the OpenCV [31] implementation of the HOG [1] algorithm. From the HOG, we retained all the detections with a strictly positive SVM score. Fig. 1 shows some intermediate results for three scenarios. Part (a) illustrates the detection of blobs in the depth images. Part (b) illustrates the performance of the HOG detector. Scenes without people, like in part (c) and (e), were used to build the background model for the depth maps. The combined use of space-time and color constraints to detect people is illustrated in (d) and (f). Tracklets are in red, and the HOG windows with top scores are shown in (d) and the foreground blobs are shown in (f).

The multiscale search of the OpenCV implementation of the HOG detector examines 34,981 candidates per image. Out of these, the HOG algorithm eliminates many false positives, depending on the threshold used on the SVM score. Adding depth information by our method improves detection performance significantly. In Fig. 2, we plot two curves for false positives (`fp`) and true positives (`tp`) for different HOG score thresholds. These curves relate results without and with the use of depth information. When the HOG score threshold is zero, our method reduces the number of `fp` from 2,594 to 76, while the number of `tp` increases from 245 to 615. When the threshold is set to 3.15, the highest value that results in some HOG detections, the number of `fp` goes from 2 to 7 and that of `tp` goes from 0 to 16. Overall, with our approach, the number of false positives is greatly reduced, and the number of true positives is simultaneously increased.

Conclusion

In this paper, we presented a strategy to combine depth-based tracking and appearance-based HOG people detection. This strategy greatly improves *both* precision and recall. Our object detector is computationally efficient and accurate. Overall, our strategy seems to give excellent results in indoor environments. In the future, we plan to explore less greedy methods for the construction of tracklets, more nuanced models of image similarity and space-time closeness, and more detailed models of sensor uncertainty. We also plan to extend our method to part-based person detection methods.

Our study suggests that the emergence of new, inexpensive depth sensors presents new opportunities for surveillance, activity analysis and people tracking. Nonetheless, these sensors are unlikely to supplant regular cameras altogether. This is because current depth sensors typically project infrared light, either temporally modulated or spatially structured, on the scene. Black or dark surfaces do not reflect well, and sometimes not at all, making background subtraction harder, and creating difficulties with people with dark hair or clothing. In addition, depth sensors are inherently limited to shorter distances because eye safety demands low illumination power levels. However, when the right conditions are met, range sensors provide an invaluable resource of information that can enhance the performance of demanding perceptual tasks.

References

1. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Computer Vision and Pattern Recognition. vol. 1, pp. 886–893 (2005)
2. Dalal, N., Triggs, B., Schmid, C.: Human Detection using Oriented Histograms of Flow and Appearance. European Conference on Computer Vision pp. 428–441 (2006)
3. Dalal, N.: INRIA Person Database. `http://pascal.inrialpes.fr/soft/olt/` (September 2010)
4. Elfes, A.: Using Occupancy Grids for Mobile Robot Perception and Navigation. Computer 22(6), 46–57 (2002)
5. Gavrila, D.: Pedestrian Detection from a Moving Vehicle. European Conference on Computer Vision pp. 37–49 (2000)
6. Gavrila, D., Giebel, J., Munder, S.: Vision-based Pedestrian Detection: The Protector System. In: Intelligent Vehicles Symposium. pp. 13–18 (2004)
7. Gavrila, D.: The Visual Analysis of Human Movement: A Survey. Computer Vision and Image Understanding 73(1), 82–98 (1999)
8. Giles, J.: Inside the Race to Hack the Kinect. The New Scientist 208(2789) (December 2010)
9. Gordon, G., Darrell, T., Harville, M., Woodfill, J.: Background Estimation and Removal based on Range and Color. In: IEEE Computer Vision and Pattern Recognition. vol. 2 (1999)
10. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling Inter-Camera Space-Time and Appearance Relationships for Tracking Across Non-Overlapping Views. Computer Vision and Image Understanding 109(2), 146–162 (2008)
11. Johansson, G.: Visual Perception of Biological Motion and a Model for its Analysis. Perceiving Events and Objects 3 (1973)
12. Kelly, M.: Visual Identification of People by Computer. Ph.D. thesis, Stanford University (1971)
13. Lowe, D.: Object Recognition from Local Scale-invariant Features. In: IEEE International Conference on Computer Vision. p. 1150 (1999)
14. Micilotta, A., Ong, E., Bowden, R.: Detection and Tracking of Humans by Probabilistic Body Part Assembly. In: British Machine Vision Conference. vol. 1, pp. 429–438 (2005)
15. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human Detection based on a Probabilistic Assembly of Robust Part Detectors. European Conference on Computer Vision pp. 69–82 (2004)
16. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based Object Detection in Images by Components. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(4), 349 (2001)
17. Muñoz, R., Aguirre, E., García, M.: People Detection and Tracking using Stereo Vision and Color. Image and Vision Computing 25(6), 995–1007 (2007)
18. Nelder, J., Mead, R.: A Simplex Method for Function Minimization. The Computer Journal 7(4), 308 (1965)
19. Papageorgiou, C., Poggio, T.: A Trainable System for Object Detection. International Journal of Computer Vision 38(1), 15–33 (2000)
20. Phillips, P.: Human Identification Technical Challenges. In: IEEE International Conference on Image Processing (2002)
21. Ramanan, D., Forsyth, D., Zisserman, A.: Strike a Pose: Tracking People by Finding Stylized Poses. In: IEEE Computer Vision and Pattern Recognition. pp. 271–278 (2005)

22. Roberts, T., McKenna, S., Ricketts, I.: Human Pose Estimation using Learnt Probabilistic Region Similarities and Partial Configurations. European Conference on Computer Vision pp. 291–303 (2004)
23. Ronfard, R., Schmid, C., Triggs, B.: Learning to Parse Pictures of People. European Conference on Computer Vision pp. 700–714 (2006)
24. Rubner, Y., Tomasi, C., Guibas, L.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 40(2), 99–121 (2000)
25. Schwartz, W., Kembhavi, A., Harwood, D., Davis, L.: Human Detection using Partial Least Squares Analysis. In: IEEE International Conference on Computer Vision. pp. 24–31 (2010)
26. Swets, J., Dawes, R., Monahan, J.: Better Decisions through Science. Scientific American p. 83 (2000)
27. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Elsevier (2009)
28. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: IEEE Computer Vision and Pattern Recognition. vol. 1 (2001)
29. Viola, P., Jones, M., Snow, D.: Detecting Pedestrians using Patterns of Motion and Appearance. International Journal of Computer Vision 63(2), 153–161 (2005)
30. Vrubel, A., Bellon, O., Silva, L.: Planar Background Elimination in Range Images: A Practical Approach. In: IEEE International Conference on Image Processing. pp. 3197–3200 (2009)
31. Willow Garage: OpenCV. `http://opencv.willowgarage.com` (September 2010)
32. Xu, F., Fujimura, K.: Human Detection using Depth and Gray Images. In: Advanced Video and Signal Based Surveillance. pp. 115–121. IEEE (2003)
33. Zhao, L., Davis, L.: Closely coupled object detection and segmentation. In: IEEE International Conference on Computer Vision. pp. 454–461 (2005)
34. Zhao, L., Thorpe, C.: Stereo and Neural Network-based Pedestrian Detection. IEEE Transactions on Intelligent Transportation Systems 1(3), 148–154 (2000)