Counting Pedestrians in Bidirectional Scenarios Using Zenithal Depth Images

Pablo Vera, Daniel Zenteno, and Joaquín Salas*

Instituto Politécnico Nacional
Cerro Blanco 141, Colinas del Cimatario, Querétaro, México, 76090
pvera@ipn.mx, ezentenoj1100@alumno.ipn.mx, jsalasr@ipn.mx

Abstract. In this document, we describe a people counting system that can precisely detect people as they are seen from a zenithal depth camera pointing at the floor. In particular, we are interested in scenarios where there are two preferred directions of motion. In our framework, we detect people using a Support Vector Machine classifier, follow their trajectory by modeling the problem of matching observations between frames as a bipartite graph, and determine the direction of their motion with a bi-directional classifier. We include experimental evidence, from four different scenarios, for each major stage of our method.

1 Introduction

Counting automatically the number of people passing a specific point is a function of paramount importance in applications such as surveillance, monitoring, and interaction between humans and machines. For instance, imagine a civil protection situation taking place in a building: People continuously enter and leave when suddenly an alarm sounds, indicating that the building must be evacuated. One can imagine how useful it is for people in charge of the evacuation procedure to query an automated monitoring system to figure out how many people are still in the building and in what areas.

In this document, we describe a people counting system that can precisely detect people as they are seen from zenithal depth cameras pointing downwards. There are a number of advantages of this configuration, including the fact that people are less likely to occlude one another and that privacy may be better protected. In particular, we are interested in scenarios where there are two preferred directions of motion, such as hallways, where people move primarily in two opposing directions, or entrances, where pedestrians pass in or out. In this paper, we introduce the detection of people in depth images using zenithal cameras. This allows us to develop a robust counting system where all entities except people are readily discarded.

In our method, we detect people, follow their movement, and determine their direction of motion. Our people detector is based on an application of Dalal and Triggs' method of the histograms of oriented gradients[5]. Then, we construct *tracklets* (chronological sequences of observations) of people by matching observations, which

* Thanks to Cole Shaw and Paul Riley for their thoughtful review of the manuscript. Also, thanks to Omar Rodríguez for installing the cameras used in this work. This research was partially supported by SIP-IPN through grant number 20121642.

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2013, LNCS 7914, pp. 84–93, 2013.

include appearance and space-time information, in a bipartite graph. Finally, tracklets are classified by modeling their normalized destination and defining a quadratic classification surface[19].

The rest of the document is organized as follows (see Fig. 1). In §2 we present a survey of the reported research related to counting people, with particular emphasis on the use of zenithal cameras. Then, as the measurement of the height will be very important in later stages of our method, in §3, we describe how we obtain a geometric description of the floor underneath the camera. Next, we describe how the people detector is implemented in §4, and how the people tracklets are constructed in §5. The direction classifier is introduced in §6. In §7, we assess the performance and show some results from the implementation of the methods described. Finally, we conclude and outline possible future applications of this research.

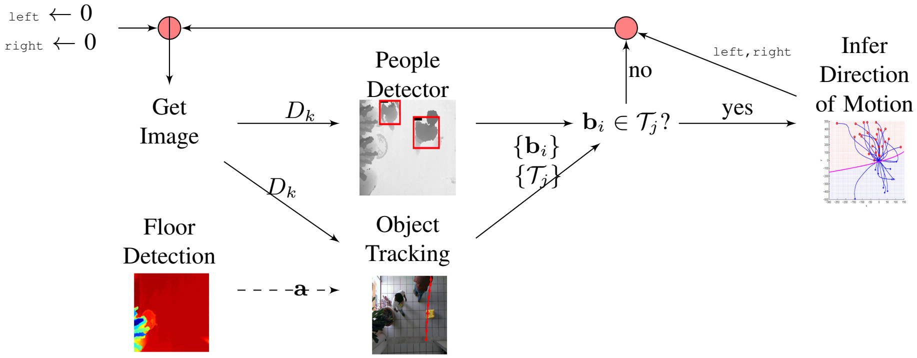


Fig. 1. Solution Global Scheme. By computing the floor parameters, **a**, it is possible to normalize the measurements of height. Thus, each time a new image D_k is read, there are two tasks to solve. One is to track the moving objects and produce tracklets $\{\mathcal{T}_j\}$. The other is to detect people, which will eventually give a set of bounding boxes $\{\mathbf{b}_i\}$. Whenever the tracklet contains a person, the counters `left` and `right` are updated depending on the direction of motion.

2 Related Work

Due to the considerable number of applications, ranging from entertainment to security, the research efforts aiming to produce reliable and fast people counters has been extensive. Nowadays, there are two main directions of research[14]: Pedestrian detection and tracking-based methods, and feature regression-based methods. In the former, pedestrians are singled out [2,24,11] and the trajectory of each individual is known. In the latter, the problem is framed as a classification one. During learning, an observed set of feature descriptors, such as edges and texture, is correlated with the number of people present. During operation, the feature descriptors are classified and the estimated number of people in the scene is obtained as a result, preserving privacy to some extent [1,12].

In our case, we focus our attention on those applications where a top view image is considered better because there are either fewer occlusions or because there are some privacy concerns that are of prime importance. For other applications where the interest is primarily frontal view images, the interested reader is referred to the reviews of Enzweiler[6] and Gavrila or Geronimo *et al.*[9], where the former places more emphasis on monocular vision and the latter on driver assistance systems. Raheja *et al.*[16] review some of the techniques commonly used in the computer vision community to count people. As a general rule, background models[22] are constructed, and the objects of interest are found via background subtraction. One key component of our approach is that people detection is done directly from depth images, extending the method proposed by Dalal and Triggs[5] to detect people in frontal-view, intensity images. However, other classification-based methods may be suitable. For instance, Haubner *et al.*[10] propose a system to detect body parts above and around a tabletop setup using a depth camera. Their work is based on the frontal view people detection framework proposed by Shotton *et al.*[20]. Importantly, the subjects are dressed with special clothes where colors signal specific portions of the body. Thus the head is covered with a red mask, the neck is yellow, the shoulders are green, the main trunk is violet, and so on. Depth images have been used previously to process zenithal images, in the form of stereo pairs. For instance, Yahiaoui *et al.*[23] detect people by dividing the observed depth map into intervals and applying a morphological operator to seek circular shapes representing the head. To our knowledge, this document reports the first counting system based on top-view depth images that uses a classifier to detect people.

3 Floor Detection

To use the height as a descriptor of the objects we observe, we normalize our measurements and express them with respect to the scenario floor, which is assumed to be flat. To that end, we rotated the 3D points computed from the depth images in order that the z-axis would coincide with the floor's normal orientation. The floor's plane is defined by the equation: $ax + by + cz = d$, where $\mathbf{x} = (x, y, z)^T$ are the coordinates of a point belonging to the plane. An appropriate scale factor can be obtained to make $(a, b, c)^T$ a unitary vector normal to the plane. Then, the rotation matrix $\mathbf{R}$ for rotating the 3D point cloud will satisfy

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \mathbf{R} \begin{bmatrix} a \\ b \\ c \end{bmatrix}. \tag{1}$$

In our case, $\mathbf{R}$ is defined by two angles, α and β, corresponding to rotation over the y and x axes, as $\begin{bmatrix} \cos\alpha & \sin\alpha\sin\beta & -\sin\alpha\cos\beta \\ 0 & \cos\beta & \sin\beta \\ \sin\alpha & -\cos\alpha\sin\beta & \cos\alpha\cos\beta \end{bmatrix}$.

Assuming that the floor either is the only plane or the plane with the largest area visible in the image, α and β can be determined as

$$\arg\max_{\alpha,\beta} N_{\max}, \tag{2}$$

where $N_{\max}$ is the highest value of frequency on a histogram of the z value of the 3D points after being rotated by $\mathbf{R}$. A large value of $N_{\max}$ indicates that the points on the floor are well aligned to have roughly the same z value.

We solved (2) using the Downhill Simplex method[15]. To avoid local minima during the minimization process, we used an initial simplex with one of the vertices having values of α and β computed with the RANSAC method[8]. This computation is done by selecting sets of 3 points at random, computing the parameters of the plane which includes each set of points, and then selecting the plane for which the maximal $N_{\max}$ is obtained, where α and β are computed using

$$\alpha = \arcsin a, \text{ and } \beta = \arcsin \frac{b}{-\cos \alpha}. \tag{3}$$

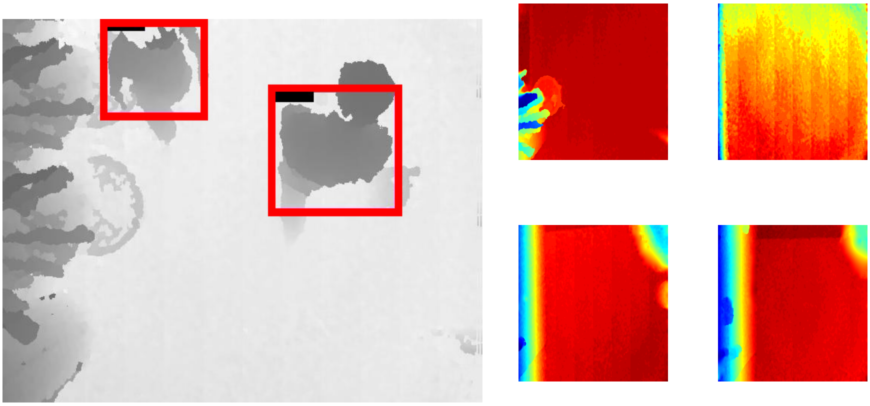
We added the Downhill Simplex optimization stage in conjunction with the RANSAC because it results in a better solution to (2). Besides, the computation complexity is low and the estimation is done offline, as the rotation matrix is computed only once after the camera is fixed.

4 People Detection

In [5], Dalal and Triggs introduced a people detector model for intensity-based images and people in upright positions. In their approach, a candidate was described in terms of features extracted from an image that is divided into same-size overlapping blocks. The blocks themselves are divided equally in same-size cells, and the amount of overlap between blocks corresponds to the size of the cell. The image features Dalal and Triggs use are based on histograms of oriented gradients extracted from the cells, weighted by the magnitude of the gradient in the block. For classification, they used a Support Vector Machine (SVM)[4]. In their method, the candidates in the image were sought using a pyramidal search.

As the intensity represents properties of the blending of many factors including the light sources, the materials of the surfaces and their geometrical structure, and the camera operating function, some researchers [21] have tested successfully a research hypothesis stating that depth information will reflect more clearly the structural properties of the objects being characterized. However, these experiments have been done on images of people taken from frontal view cameras. In the problem of counting the number of people that pass by, frontal view images may result in occlusions that could compromise the objective. In what follows, we describe how we have adapted the Dalal and Triggs methodology to detect people using zenithal depth cameras.

We define a standard size bounding box for people detection of 96×96 `pixels`. The size was chosen by calculating an average over 993 true positive samples and approaching the resulting size to suitable cell and block elements. The sample selection for training is illustrated in Figure 2. As in [5], the features were extracted by computing the gradient over cell structures, which corresponds to neighborhoods of 8×8 pixels. The orientation of the gradient was clustered into nine-bin histograms. The frequency was weighted using the magnitude of the gradient blocks, of 2×2 cells. That is, a person is described by a feature vector $\mathbf{x}$ of size 1089×1.



(a) Positive sample

(b) Scenarios without people

Fig. 2. Creating samples for training the people detector. Illustration of some positive samples and scenarios from where the negative samples were obtained to train the classifier. Samples were created by selecting manually or randomly bounding boxes on scenarios with people and without people, respectively, by visual inspection of the images.

5 Tracking

As activity develops below the camera, some objects are detected in the scene. Let $\{O_{j,1},\ldots,O_{j,m}\}$ correspond to the m objects detected during frame j. In our case, we are interested in inferring where the objects that were detected during frame i have moved in frame j. This tracking problem has been the subject of a large number of papers (for a recent review, please see[18]). We are particularly interested in the problem where the inference process occurs between consecutive frames, and where, as consequence of the observations made, we have appearance and temporal-location information. That is, let a particular observation $O_{j,a} = (h_{j,a}, \mathbf{x}_{j,a}, t_{j,a})$ consist of the location $\mathbf{x}_{j,a}$ and time $t_{j,a}$ where its maximum height $h_{j,a}$ was observed. Just as in a number of other articles[3,17], we model the problem of matching observations in frame j with observations in frame $j+1$ as a complete bipartite weighted graph $G = (O, E, P)$. The vertices $O = \{O_j, O_{j+1}\}$ are divided between the observations made in the frame j and the one that follows it. The set of edges $E = \{e_{j,a}^{j+1,b}\}$ represents the hypothesis that two particular observations correspond to the same person. And the weight $P = \{p_{j,a}^{j+1,b}\}$ represents the likelihood of a particular hypothesis about the correspondence of two observations.

In our approach, we express the weight between two observations as

$$p_{j,a}^{j+1,b} = 1 - p(h_{j,a}, h_{j+1,b}) \cdot p(\mathbf{v}_{j,a}^{j+1,b}) \cdot p(t_{j,a}, t_{j+1^b}), \qquad (4)$$

where $p(h_{j,a}, h_{j+1,b})$ is defined as

$$p(h_{j,a}, h_{j+1,b}) = e^{-\alpha|h_{j,a}-h_{j+1,b}|}, \qquad (5)$$

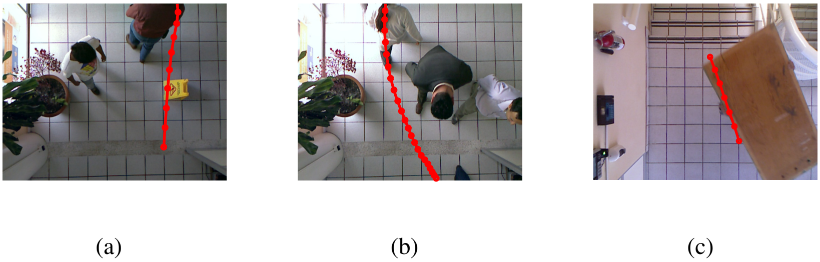

(a) (b) (c)

Fig. 3. Trackelet construction. The dots in the line correspond to the highest point positions. The illustration is done in color images but the process was done with the depth images. This simple strategy fails to count pedestrians when there are no people, as in (c). Therefore, there is a need to couple it with a people detector.

and α represents the weight for how likely it is to observe heights h_j^a and h_{j+1}^b; correspondingly $p(\mathbf{v}_{j,a}^{j+1,b})$ is defined as

$$p(\mathbf{v}_{j,a}^{j+1,b}) = e^{-\beta\left|\frac{\mathbf{x}_{j,a}-\mathbf{x}_{j+1,b}}{t_{j,a}-t_{j+1,b}}\right|^n}, \tag{6}$$

and β is the weight associated with observing a certain velocity $\mathbf{v}_j^{a,b}$ between two observations; then, $p(t_{j,a}, t_{j+1^b})$ is defined as

$$p(t_{j,a}, t_{j+1,b}) = e^{-\gamma|t_{j,a}-t_{j+1,b}|^n}. \tag{7}$$

Here, γ is the weight corresponding to observing the same object at times $t_{j,a}$ and t_{j+1}^b. Note that n in (6) and (7) has the purpose of establishing a threshold on the observable average velocity and time values. That is, the larger the value of n, the steeper it is the slope of (6) and (7) around the value of β and γ, respectively. In practice, this has the effect of accepting low values but being critical for values larger than β or γ. Also, note that the third term in (4) is introduced to avoid accepting observations that spread in time.

Now, the problem is to find a match of observations corresponding to consecutive frames at minimal cost. Although other more sophisticated methods could be used, given the usually reduced size of the graphs, we have used the Hungarian algorithm[13]. Illustrations of actual tracklets are presented in Fig. 3.

6 Bi-directional Classifier

In our method, a tracklet $\mathcal{T}^k = \{\mathbf{t}_{j+1}^k, \ldots, \mathbf{t}_{j+m}^k\}$ corresponds to a person when the detector responds positively to a certain number of observations within the tracklet. Given $\mathbf{b}$ as the center of the person's detected bounding box, and $\mathbf{t}_{j+n}^k$ as the position of the tracked object, the observation is said to correspond to the detection whenever $\left\|\mathbf{t}_{j+n}^k - \mathbf{b}\right\| < \tau$, for a predefined value of τ.

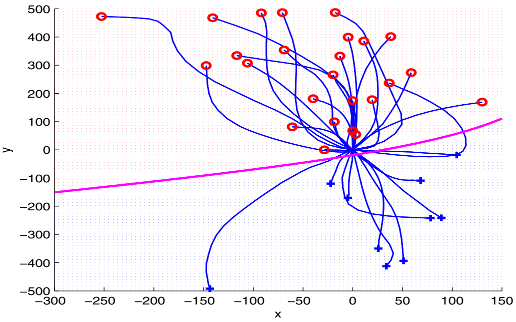


Fig. 4. Classification of the direction of motion. The trajectories are normalized to start at $(0,0)$. The end point is circled or crossed for the trajectories going in opposite directions. The curved line crossing horizontally corresponds to the decision surface.

In our problem, we want to classify a person's trajectory as going in one of two opposite directions. Thus, given a trajectory $\mathcal{T}^k$, we want to classify it in either one of two disjoint sets $\mathcal{T}_l$ and $\mathcal{T}_r$, which represent opposite directions. To that end, we define centered trajectories as $\overline{\mathcal{T}}^k = \{\overline{\mathbf{t}}^k_{j+1}, \ldots, \overline{\mathbf{t}}^k_{j+m}\}$, where $\overline{\mathbf{t}}^k_{j+n} = \mathbf{t}^k_{j+n} - \mathbf{t}^k_{j+1}$, for $n = 1, \ldots, m$. The problem now is to define a surface to distinguish between classes. Although more sophisticated methods can be applied, in our approach we fitted multivariate normal densities with covariance estimates computed by class, resulting in a decision surface of the form $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{b} + c$,

$$D(\overline{\mathcal{T}}^k) = \begin{cases} \mathcal{T}_l & f(\overline{\mathbf{t}}^k_{j+m}) < 0, \\ \mathcal{T}_r & \textit{otherwise}, \end{cases} \tag{8}$$

where $\mathbf{A}$, $\mathbf{b}$, and c correspond to the quadratic, linear, and constant coefficients computed as in [19].

7 Experimental Results

We implemented and tested the methods described in this article in four different scenarios. To obtain the depth images, we used four Microsoft Kinects. To grab our images and process the results we used a variety of software tools including OpenCV, OpenKinect, Linux, Matlab, and Eclipse.

For the people detector, we collected 993 positive images and 1,038 negative images, of which 20% were used for testing and the rest for training the SVM classifier (see Figure 2). To assess the performance of the classifier, we constructed a *precision-recall* curve where the SVM margin was varied between -1 and 1. Note that precision is related to the fraction of detections that are correct, whereas the recall is the fraction of detections out of the possible detections. The result is shown in Fig. 5. As customary, we follow the practice at the Pascal Challenge[7], where a detection is declared when the following condition is met

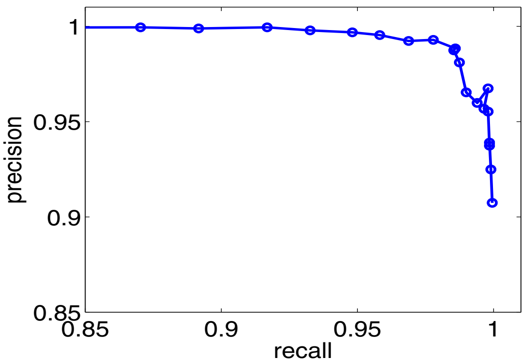


Fig. 5. Precision-recall curve for the thresholds between -1 and 1, in intervals of 0.1. The circles between the segments correspond to the evaluations.

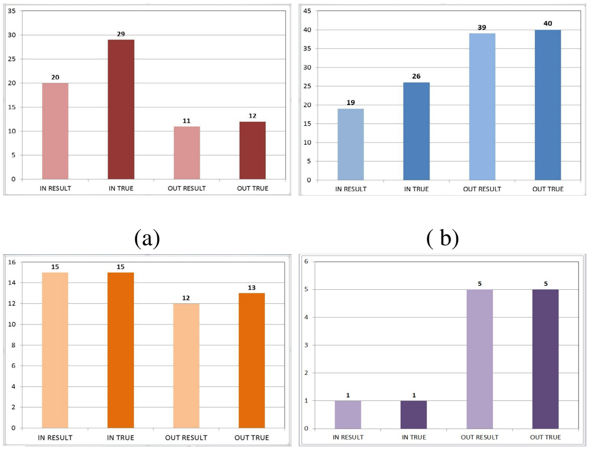
$$t_p = \begin{cases} 1 \; \mathit{if} \, \frac{area(b_d \cap b_g)}{area(b_d \cup b_g)} > 0.5, \\ 0 \qquad \mathit{otherwise}, \end{cases} \tag{9}$$

where b_d and b_g correspond to the bounding box of the detection and the ground truth (i.e., test samples labeled as positives or negatives), respectively. This curve is very useful to find a suitable threshold for the SVM margin that satisfies a commitment between sensitivity and a true positive rate or specificity of detections.

During tracking, we extracted foreground objects by subtracting the current depth image from a depth image without objects. Then, we found the pixel corresponding to the highest point in the depth map, and discarded all the pixels corresponding to the connected component for that pixel. This procedure was repeated until there was no point that was 0.2 ($\mathtt{m}$) above the floor. The values of the constants in (5) and (6), and the values of α, β and n are 1, 0.8, and 6 respectively.

To generate the bi-directional classifier we used a sequence with 1,000 (`frames`) for each scenario. After a set of tracklets was obtained, each tracklet was normalized to make it coincide with the origin. Each side was modeled with a multivariate normal density with covariance estimates computed by class. For instance, for the curve corresponding to the scenario illustrated in Fig. 4 the coefficients for $\mathbf{A}$, $\mathbf{b}$ and c correspond to $10^{-4}\begin{pmatrix} 0.2579 & -0.5389 \\ -0.5389 & 0.016 \end{pmatrix}$, $\begin{pmatrix} -0.0178 \\ 0.03 \end{pmatrix}$, and 0.5249, respectively.

To count the number of people, we select a one hour sequence of images for each scenario. In Fig. 6 we show the results organized as pedestrians going in one direction (*in*) or the opposite direction(*out*). We show two columns corresponding to the result of our system and ground truth obtained by visual inspection of the sequences. The results vary widely but are nevertheless encouraging. Comparing results with other methods is difficult as they may vary depending on the particular sequence being studied. From our work, our sequences are available for review at `http://imagenes.cicataqro.ipn.mx/CountingPedestrians/`.



(a) (b)

(c) (d)

Fig. 6. Four different scenarios were tested collecting images for an hour. The bars show people going in one direction (*in*) or another (*out*). For each direction, we show the result of our method and ground truth.

8 Conclusion

In this document, we introduced a computer vision system to count the number of pedestrians in a bidirectional scenario. In our approach, we have introduced a people detector based on depth images (inspired on the scheme proposed by Dalal and TriggsTriggs[5]) that allows us to deal with complex scenarios in which what is moving in front of the camera may not necessarily be a person. We have included experimental evidence for each major stage of out method. In particular, as for the number of pedestrians that are being counted, the results, although varying with the scenario, are nevertheless encouraging.

In the future, we plan to streamline every stage of the processing sequence and experiment with more sophisticated scenes, such as those resulting from evacuations, where crowds of pedestrians move very close to each other but where the advantages of a zenithal camera to reduce the problems associated with occlusions can be demonstrated fully.

References

1. Chan, A., Liang, Z., Vasconcelos, N.: Privacy Preserving Crowd Monitoring: Counting People without People Models or Tracking. In: CVPR, pp. 1–7 (2008)
2. Chen, T., Chen, T., Chen, Z.: An Intelligent People-flow Counting Method for Passing through a Gate. In: Conference on Robotics, Automation and Mechatronics, pp. 1–6 (2006)
3. Chowdhury, A., Chatterjee, R., Ghosh, M., Ray, N.: Cell Tracking in Video Microscopy using Bipartite Graph Matching. In: ICPR, pp. 2456–2459. IEEE (2010)

4. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines: And other Kernel-based Learning Methods. Cambridge University Press (2000)
5. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR, pp. 886–893 (2005)
6. Enzweiler, M., Gavrila, D.: Monocular Pedestrian Detection: Survey and Experiments. IEEE Trans. on Pattern Anal. and Mach. Intell. 31(12), 2179–2195 (2009)
7. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) Challenge. Int. J. Comput. Vision 88(2), 303–338 (2010)
8. Fischler, M., Bolles, R.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the ACM 24(6), 381–395 (1981)
9. Geronimo, D., Lopez, A., Sappa, A., Graf, T.: Survey of Pedestrian Detection for Advanced Driver Assistance Systems. IEEE Trans. on Pattern Anal. and Mach. Intell. 32(7), 1239–1258 (2010)
10. Haubner, N., Schwanecke, U., Dorner, R., Lehmann, S., Luderschmidt, J.: Towards a Top-View Detection of Body Parts in an Interactive Tabletop Environment. In: Architectures of Computing Systems (2012)
11. Kilambi, P., Ribnick, E., Joshi, A., Masoud, O., Papanikolopoulos, N.: Estimating Pedestrian Counts in Groups. Computer Vision and Image Understanding 110(1), 43–59 (2008)
12. Kong, D., Gray, D., Tao, H.: A Viewpoint Invariant Approach for Crowd Counting. In: ICPR, vol. 3, pp. 1187–1190 (2006)
13. Kuhn, H.: The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly 2(1-2), 83–97 (2006)
14. Li, J., Huang, L., Liu, C.: Robust People Counting in Video Surveillance: Dataset and System. In: AVSS, pp. 54–59 (2011)
15. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes. Cambridge University Press (2007)
16. Raheja, J., Dutta, P., Kalita, S., Lovendra, S.: An Insight into the Algorithms on Real-Time People Tracking and Counting System. Int. J. of Comp. Appl. 46(5), 1–6 (2012)
17. Rowan, M., Maire, F.D.: An Efficient Multiple Object Vision Tracking System using Bipartite Graph Matching. In: FIRA Robot World Congress. FIRA Robot World Congress (2004)
18. Salti, S., Cavallaro, A., Di Stefano, L.: Adaptive Appearance Modeling for Video Tracking: Survey and Evaluation. IEEE Trans. on Image Process. (2012)
19. Seber, G.: Multivariate observations. Wiley (1984)
20. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR (2011)
21. Spinello, L., Arras, K.: People Detection in RGB-D Data. In: IROS, pp. 3838–3843 (2011)
22. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: Principles and Practice of Background Maintenance. In: ICCV, p. 255 (1999)
23. Yahiaoui, T., Meurie, C., Khoudour, L., Cabestaing, F.: A People Counting System based on Dense and Close Stereovision. In: Image and Signal Processing, pp. 59–66 (2008)
24. Zhang, E., Chen, F.: A Fast and Robust People Counting Method in Video Surveillance. In: International Conference on Computational Intelligence and Security, pp. 339–343 (2007)