People Counting based on Kinect Depth Data

Rabah Iguernaissi, Djamal Merad, Pierre Drap

People Counting based on Kinect Depth Data

Rabah Iguernaissi, Djamal Merad and Pierre Drap
Aix-Marseille University, LSIS - UMR CNRS 7296, 163 Avenue of Luminy, 13288 Cedex 9, Marseille, France

Keywords: People Counting, Depth Data, Intelligent Sensor, People Detection.

Abstract: The people's counting is one of the most important parts in the design of any system for behavioral analysis. It is used to measure and manage people's flow within zones with restricted attendance. In this work, we propose a counting strategy for counting the number of people entering and leaving a given closed area. Our counting method is based on the use of depth map obtained from a Kinect sensor installed in a zenithal position with respect to the motion direction. It is used to count the number of people crossing a virtual line of interest (LOI). The proposed method is based on the use of two main modules a people detection module that is used to detect individuals crossing the LOI and a tracking module that is used to track detected individuals to determine the direction of their motions. The people detection is based on the design of a smart sensor that is used with both the grayscale image that represents depth changes and the binary image that represents foreground objects within the depth map to detect and localize individuals. Then, these individuals are tracked by the second module to determine the direction of their motions.

1 INTRODUCTION

The study of people's behavior is one of the most important purposes in the design of several computer vision applications. The design of such systems is generally motivated by either the automatization or the improvement of some procedures that relay on human operators. One of the most studied aspects is the design of automatic counting systems that are able to either count the number of individuals present within a region of interest (ROI systems) or the number of individuals crossing a line of interest (LOI systems).

People's counting systems are used to obtain information about people flow in closed areas such as stores, metros, and malls. These information may be used for several purposes either statistical to improve the monitoring of some areas such as stores by adapting the stuff during peak hours or for safety purpose in surfaces with limited number of attendance by producing a count that gives an alert whenever the maximal permitted number of individuals is reached.

In the last few years, many applications that integrate computer vision techniques were proposed to realize an automatic counting system that is able to count the number of individuals crossing a line of interest based on the use of either classical RGB cameras or depth sensors. In most works, the sensors are installed in a zenithal position with respect to the motion direction to reduce the occurrence of occlusions.

In this work, our objective is the design of a counting system that may be used to determine the number of customers entering or leaving a sale area. For this, we proposed a system that uses depth map from a Kinect sensor to count the number of individuals crossing a virtual line of interest within the Kinect field of view and determine the direction of their motion to decide whether detected individuals are entering or leaving the store.

The rest of this paper is organized in three main parts. The first part is dedicated to the study of some related works that were published in this field. In the second part, our counting strategy is presented and the last section is dedicated to the presentation of some results. This paper ends up with a conclusion.

2 RELATED WORKS

The counting of people is the basic part of any system that is designed for crowd monitoring and attendance control either for statistical, marketing, or security purposes. Due to its importance many researches were done in this field. The main objective for all these works is the proposal of algorithms that enable the automatic counting of people either crossing a virtual line or determining the number of people who are present in the studied scene at given time.

Iguernaissi, R., Merad, D. and Drap, P.
People Counting based on Kinect Depth Data.
DOI: 10.5220/0006585703640370
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 364-370
ISBN: 978-989-758-276-9

Most proposed methods for pedestrians counting are based on two main steps. The first step is generally the people detection step which consists in detecting moving objects and separating them into single persons. This is done based on classical algorithms for motion detection such as background subtraction and a convenient segmentation strategy such as connected component analysis, K means,.... The second step is the tracking step in which detected individuals are tracked either to determine the direction of motion (entering or leaving a given surface) or just to avoid recounting the same individual several time while he is moving within the studied scene.

Depending on the counting objective the proposed methods for people counting can be classified into two main categories: The LOI counting methods and the ROI counting methods. The LOI methods are designed to estimate the number of people crossing a virtual Line Of Interest (LOI) within the studied scene. For this category we can mention (Ma and Chan, 2013), (Del Pizzo et al., 2016), and (Cong et al., 2009). The second category which are the ROI methods are designed for crowd estimation to estimate the number of people present within a Region Of Interest (ROI) in the studied scene at given time. For methods adopting this approach we may cite (Antić et al., 2009), (Bondi et al., 2014).

One of the most used approaches for people's counting is based on motion analysis across a virtual LOI to detect and count persons. In this approach, for counting the number of people crossing a virtual LOI in video stream Ma and Chan (Ma and Chan, 2013) proposed a method that is based on integer programming. The algorithm is designed to estimate the instantaneous people's count using local-level features and regression. The proposed method is based on constructing an image slice from the temporal slice image where each column in the slice image corresponds to the line of interest at given time t. Then, the resulting region of interest is studied to detect blobs that corresponds to crowd segment crossing the LOI using local HOG features. Finally, the count of people in each crowd segment is estimated by Bayesian Poisson regression. In contrast to this Cong et al. (Cong et al., 2009) designed a method that is based on flow velocity field estimation. In this algorithm the first step consists in detecting the velocity field on the LOI which is segmented according to the moving direction. Then, a dynamic mosaic is used to construct blobs that in their turn will be used to estimate the number of people based on regression on their area and number of edges.

Another common approach consists in detecting moving individuals or a part of their bodies (generally heads) in either classical RGB cameras or depth sensors. Then, tracking these individuals to count people crossing a virtual LOI or people present within a ROI. Most of methods in this category starts by detecting moving objects in the studied scene based on classical methods used for motion detection such as the background subtraction (Antić et al., 2009) and (Bondi et al., 2014) or the frame differencing in (Chen et al., 2012). The second step consists in segmenting the detected blobs to detected individual (Antić et al., 2009) and (Chen et al., 2012) or their heads such as in (Fu et al., 2014), (Bondi et al., 2014), (Van Oosterhout et al., 2011), and (Zhang et al., 2012).

In this context, Antic et al. (Antić et al., 2009) proposed a counting method in video stream. This method is based on three basic steps. The first step consists in the detection of foreground object using a classical background subtraction method. In the second step, the foreground is clustered by K-means into the maximum possible number of clusters which is supposed to be the number of persons who are present in the studied ROI. The third and last step consists in the tracking of detected individuals from the previous step based on a greedy solution to a dynamic assignment problem between clusters in consecutive frames. Then, the end points of tracks are used to increment the entrance or the exit counter. In the same manner, Chen et al. (Chen et al., 2012) proposed a method in which the motion detection is done by frame differencing and the blobs segmentation to single individuals is done by connected component analysis. Then, tracking is done by bounding boxes intersection-check technique by supposing that people in crowd are moving slowly.

For the same purpose many methods were proposed to detect heads of individuals who are present in the studied ROI. Most of these algorithms exploit the fact that a head is the closest part of human body to a ceiling depth sensor that is installed in a zenithal position with respect to the studied scene. Among these methods, we may mention Zhang et al. (Zhang et al., 2012) who proposed a method based on a water filling algorithm. The proposed method consists in the detection of local minimums of water drops needed to fill a depth map which is obtained by a Kinect sensor. These local minimums are supposed to be heads as they are the closest parts to the Kinect sensor which is installed in a zenithal position with respect to the ROI. In the same context, Bondi et al. (Bondi et al., 2014) used depth data from a stereo system to localize heads in the studied scene by localizing the local minimum in each detected blob representing a foreground object. These foreground blobs are obtained by associating a classical background subtraction method

with a simple edge detection. The association of this head detection technique with a simple tracking strategy allow the counting of people who are present in the ROI.

Another category of methods that are dealing with the problem of pedestrian's counting are methods based on the design of smart sensor that interprets data from an RGB camera or depth sensor to count people within a ROI. For this category, Del Pizzo et al. (Del Pizzo et al., 2015) proposed a method for people counting that is based on two main steps. The first step consists in simple background subtraction method that is used to detect foreground objects in the studied scene. Then, the second step consist in people counting based on the use of smart sensor that is able to interpret the results of foreground detection. The used sensor has a rectangular shape which is divided into several cells that are activated whenever the number of foreground pixels within the cell is larger than a threshold. The evaluation of the activation sequence allows the detection and the counting of people moving within the ROI. In (Del Pizzo et al., 2016), the same basic idea of using rectangular sensor that is divided into several cells was adopted to detect people. This basic idea is associated to a method similar to the working of incremental rotary encodes (Billingsley et al., 2008) to detect the motion direction. The evaluation of the activation sequence associated to the direction of motion enables the counting of people crossing a virtual LOI. In the same context, Barandiaran et al. (Barandiaran et al., 2008) proposed a method that is based on the use of a fixed number of equidistant virtual lines that are placed orthogonally to the expected motion direction in the ROI. Then, the people counting for each virtual line is done by detecting the non-zero intervals in the difference image between current and previous frame at the level of the virtual line.

The study of the state of the art gave us a general idea about the methods adopted for counting people either crossing a LOI or present within a ROI at a given time. Most recent methods are exploiting data from either an RGB camera or depth sensor that are installed in a zenithal position with respect to the studied scene. This tendency is due to the fact that the zenithal position of the used sensor reduces drastically the recurrence of occlusions which is one of the major issues in the field. But, in the other hand, the segmentation and tracking problems persist as the zenithal position of camera reduces the data that may be used to differentiate different individuals.

In this paper, we adopted a method that is based on exploiting depth data from a Kinect sensor that is installed in a zenithal position with respect to the studied scene. The use of Kinect in zenithal position reduces the recurrence of occlusion and the use of depth data for people detection improve the segmentation.

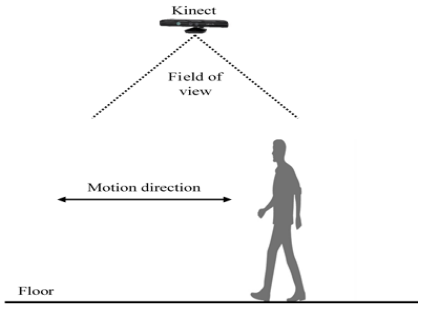

Figure 1: Sensor positioning.

The main contributions of this work are:

- The use of Kinect's depth data with a smart sensor to detect individuals crossing the LOI.
- The use of depth data from a kinect that is installed in zenithal position to construct both a binary and a grayscale images to facilitate the detection and tracking processes.

3 PROPOSED METHOD

The objective of this work is the proposal of a solution that allows counting the number of people entering and leaving a closed area. To achieve this objective, we proposed a method for counting individuals crossing a LOI based on the use of depth map obtained from a Kinect sensor. This counting approach is proposed in the context of behavioral marketing analysis to measure the flux of customers in a store.

As said previously, the people counting problem was addressed in different ways either based on classical RGB cameras or depth sensors. Most proposed methods are dealing with the same major issues which are occlusions, crowd management, and segmentation problems.

In order to deal with all these issues, we proposed a counting strategy that is based on exploiting depth data obtained from a Kinect sensor placed in a zenithal position with respect to the motion direction as shown in Figure 1. The zenithal position of the sensor reduces drastically the recurrence of occlusions even in dense crowds and the use of depth data improves segmentation in case of crowded scenes.

The proposed counting method is composed of two main modules. The first module is the people detection module which is based on exploiting depth data from a Kinect sensor to detect individuals crossing the virtual line of interest. The second module is

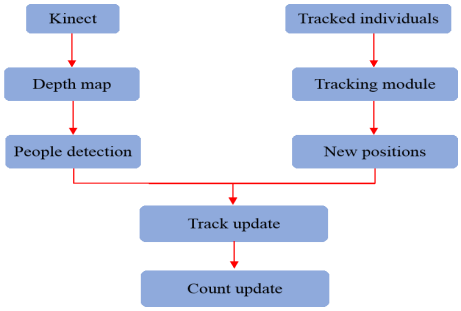

Figure 2: General diagram of the proposed method.

the tracking module that is based on the use of dedicated particle filters to track detected individuals from the first module in order to both determine the direction of their motions and avoid recounting the same individual several times while he is moving within the Kinect field of view. The proposed algorithm is illustrated in Figure 2.

3.1 People Detection Module

As said previously, the first part in our counting strategy consists in people detection based on the use of depth map obtained from a ceiling Kinect sensor. The people detection is done in several steps as illustrated in the diagram of Figure 3.

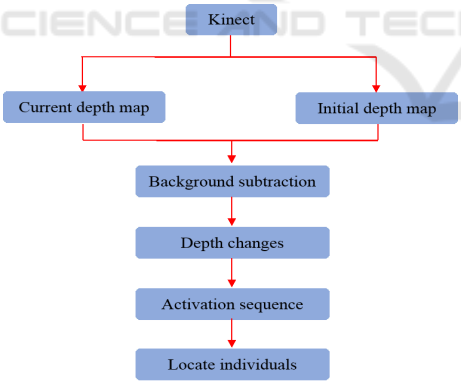

Figure 3: People detection.

The first step in this people detection scheme is the detection of depth changes that represent the foreground or the moving parts within the studied scene. For this, we used a method of background subtraction with depth data obtained from the Kinect. The background subtraction is done by frame differencing between the smoothed current and initial depth maps. The initial depth map is obtained by estimating the distance between the sensor and the non-moving objects in the scene. This initial depth map is used as background for background subtraction. In our method the background is estimated and updated based the use of running average over a patch of N frames using equation 1.

$$background = \frac{1}{N}\sum_{i=1}^{N} depthMap_i. \quad (1)$$

Where: $depthMap_i$ is the depth map at frame i and N is the total number of frames.

Ones the background estimated, both the background and the current depth maps are smoothed by applying a Gaussian filter to reduce the noise within the depth maps. The use of an adequate filter allows both removing noise and conserving the general shape of data. The resulting depth maps are used to estimate both a grayscale image D_t and a binary image B_t. The D_t image represents depth changes at time t and B_t image represents the foreground parts of the scene at the same instant t. These images are estimated from the difference image $diff_t$ that corresponds to the difference between the current and initial depth maps. The $diff_t$ is calculated based on equation 2.

$$diff_t = |depthMap_t - background|. \quad (2)$$

Where: $depthMap_t$ and $background$ are the current and initial depth maps respectively.

Ones the $diff_t$ image is calculated, it is used to create both the D_t by considering that each centimeter in the depth corresponds to a gray level in the D_t image (this assumption is based on the fact that the height of a person doesn't exceed $255cm$) and the B_t image that represents foreground objects. This image is created by supposing that depth values in $diff_t$ image that are within a certain range (th_{min} and th_{max}) are part of the foreground. These two images are created based on equations 3 and 4.

$$D_t(x,y) = \begin{cases} diff_t(x,y) & if\ diff_t(x,y) < 256 \\ 0 & otherwise \end{cases} \quad (3)$$

$$B_t(x,y) = \begin{cases} 1 & if\ \ th_{min} < diff_t(x,y) < th_{max} \\ 0 & otherwise \end{cases} \quad (4)$$

The next step consists on the design of a smart sensor that will be used to detect and segment foreground objects to both separate foreground blobs into different individuals and localize the heads of these individuals. The proposed smart sensor is designed as a rectangle that is placed over the LOI as shown in Figure 4. This rectangle has a width that corresponds to twice the estimated average of the width of the surface occupied by a single individual in the scene and

the length of the sensor corresponds to the size of the image. This rectangle is divided into several square cells with a width that corresponds to one third the estimated average width of the surface occupied by a single individual in the scene.

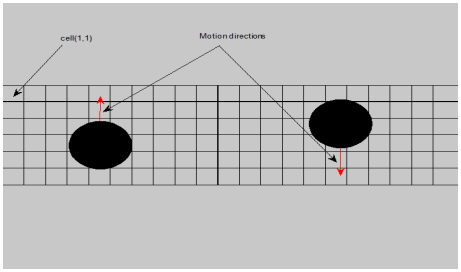

Figure 4: Design of the smart sensor.

Ones the smart sensor is designed, the next step consists in finding the activation sequence that will be used to detect and localize individuals within the scene. The recovery of the activation sequence is done in two steps. In the first step, the foreground image B_t is used to detect active cells based on the number of foreground pixels within each cell. Then if the number of foreground pixels is greater than a threshold τ the cell is active (equation 5).

$$cell_{ij} = \begin{cases} 1\ (active) & if\ \ N > \tau \\ 0\ (notactive) & otherwise \end{cases} \quad (5)$$

This step allows the detection of significant foreground parts that may present either a single or a group of individuals. Ones individuals and group of individuals are detected, the second step consists in using the grayscale image D_t to localize individuals and segment group of individuals into single individuals. This is done by detecting local maximums within the active cells. The detection of local maximums is done by summing up all pixels within each active cell based on the use of equation 6.

$$cell(i,j) = \begin{cases} \sum_{cell_{ij}} D_t & if\ \ cell_{ij}\ \ active \\ 0 & otherwise \end{cases} \quad (6)$$

Then a cell is considered to be a local maximum if the value $cell(i,j)$ is greater than all surrounding cells $(cell(i-1,j-1), cell(i-1,j), ...cell(i+1,j+1))$. The detected maximums are considered to be heads and the number of these maximums corresponds to the number of individuals crossing the LOI at time t.

3.2 Tracking Module

The second part of our counting strategy consists in the integration of a tracking module that enable tracking individuals detected in the first module to both determine the motion direction and avoid recounting the same individual several times while he is moving within the Kinect field of view.

For this, we used a tracking strategy based on the use of a dedicated particle filter similar to the one used in (Nummiaro et al., 2002) for each detected individual. The tracking is performed in the grayscale image D_t that represents depth changes. Then, each detected individual is represented by a square centered at the point (x,y) with width w and the new position for each individual is estimated based on the motion model represented by equations 7 and 8.

$$(x,y)_{\mathrm{t}} = (x,y)_{(\mathrm{t}-1)} + (u,v)_{(\mathrm{t}-1)} \bullet \Delta t \quad (7)$$

$$(u,v)_{\mathrm{t}} = (u,v)_{(\mathrm{t}-1)} \quad (8)$$

where: (x,y) and (u,v) are the position and velocity of tracked individual along the x-y axis.

Then, these new particles are weighted based on the difference between their intensity histogram and the intensity histogram of the tracked individual. The new position of the individual is determined from the position of the weighted particles according to equation 9.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \sum_{n=1}^{N} w_t^{(n)} \begin{bmatrix} x \\ y \end{bmatrix}^{(n)} \quad (9)$$

Where:(x,y) is the estimated position for the tracked individual, $w_t^{(n)}$ is the weight of the particle located at position $(x,y)^{(n)}$, and N corresponds to the number of particles used in our particle filter.

This module allow the recovery of the trajectory of tracked individuals within the Kinect field of view. These trajectories indicate the direction of motion for each individual (entering or leaving).

4 EVALUATION

For evaluating our counting strategy, we used a Kinect for windows version 1 that was installed in zenithal position with respect to the motion direction. The experiment was done in a lab. During experiment 5 individuals were recorded while moving through the Kinect field of view. These individuals were recorded in different scenarios such as "walking" and "crossing". In the first scenario "walking" the individuals were moving either close or far from each other in the same direction whereas in the second scenario "crossing" the individuals were moving in different directions (some entering and the others leaving the room).

These data were first used to evaluate our people detection method. For this, we start by estimating the background based on the use of the first 100 frames. Then, this background is used to estimate the grayscale images D_t and the binary images B_t by taking ($th_{min} = 50$ and $th_{max} = 255$). These images are used with a sensor with a cell width of 50 pixels. The results of this detection method are illustrated in Figure 5. Then, detections are used as initialization for the tracking module and the tracking results are illustrated in Figure 6.

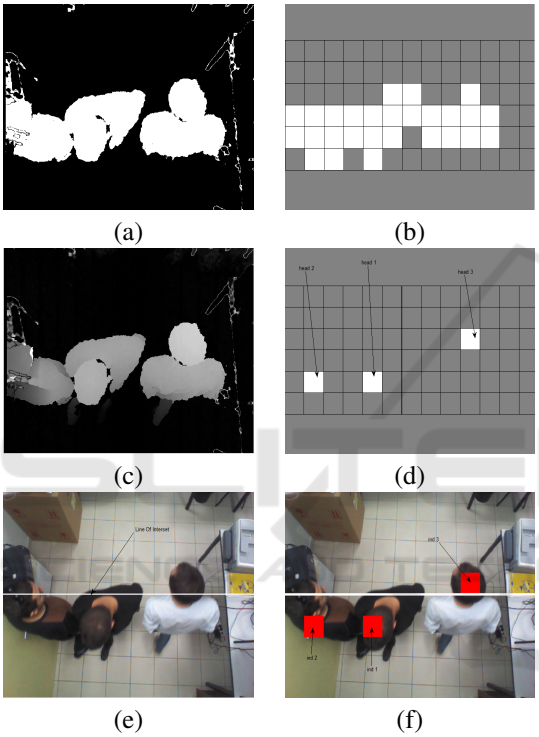

Figure 5: Detection results: (a) binary image, (b) activation sequence, (c) grayscale image , (d) people localization, (e) corresponding RGB image, (f) detected individuals.

Figure 6: Tracking results: (a) Frame 726, (b) frame 740.

From the above figures, we notice that our detection methods is able to detect individuals and segment group of individuals into single individuals even when people are close to each others. This is due to the use of depth data with a smart sensor that enable a simple detection for local maximums that represents individual's heads. Then, the use of the tracking module enable the detection of the motion direction as shown in the Figure 6.

The use of both the detection and the tracking modules allows the count of people entering and leaving the given closed area. The counting results based on the motion direction compared to the ground truth of the above experiment are shown in the diagram of Figure 7.

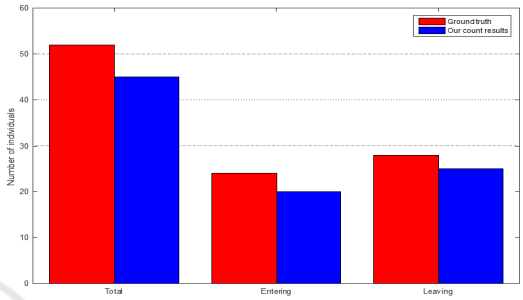

Figure 7: Counting results.

5 CONCLUSION

In this paper, we proposed a method for counting the number of people entering and leaving a closed area. The proposed strategy is based on exploiting depth data obtained from a Kinect sensor. The sensor is installed in zenithal position with respect to the motion direction in order to reduce the occurrence of occlusions.

The proposed method is based on the design of two main modules. The first module is the people detection module which is based on the use of smart sensor in both a grayscale image D_t that represents depth changes within the studied scene and a binary image B_t that represents foreground objects. The use of a smart sensor in these two images allows a good segmentation of crowds and a good localization of individuals. Ones people are detected, the second module which is the tracking module is used to track these individuals in the D_t image. The tracking module is based on the use of dedicated particle filter for each individual. The use of this module allows both determining the direction of motion (entering or leaving the store) and reducing the redundancy of count by avoiding the recount of the same individual several times while he is moving within the Kinect field of view.

REFERENCES

Antić, B., Letić, D., Ćulibrk, D., and Crnojević, V. (2009). K-means based segmentation for real-time zenithal people counting. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 2565–2568. IEEE.

Barandiaran, J., Murguia, B., and Boto, F. (2008). Real-time people counting using multiple lines. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*, pages 159–162. IEEE.

Billingsley, J., Ellin, A., and Dolsak, G. (2008). The design and application of rotary encoders. *Sensor Review*, 28(2):150–158.

Bondi, E., Seidenari, L., Bagdanov, A. D., and Del Bimbo, A. (2014). Real-time people counting from depth imagery of crowded environments. In *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, pages 337–342. IEEE.

Chen, C.-H., Chen, T.-Y., Wang, D.-J., and Chen, T.-J. (2012). A cost-effective people-counter for a crowd of moving people based on two-stage segmentation. *Journal of Information Hiding and Multimedia Signal Processing*, 3(1):12–25.

Cong, Y., Gong, H., Zhu, S.-C., and Tang, Y. (2009). Flow mosaicking: Real-time pedestrian counting without scene-specific learning. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1093–1100. IEEE.

Del Pizzo, L., Foggia, P., Greco, A., Percannella, G., and Vento, M. (2015). A versatile and effective method for counting people on either rgb or depth overhead cameras. In *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*, pages 1–6. IEEE.

Del Pizzo, L., Foggia, P., Greco, A., Percannella, G., and Vento, M. (2016). Counting people by rgb or depth overhead cameras. *Pattern Recognition Letters*, 81:41–50.

Fu, H., Ma, H., and Xiao, H. (2014). Scene-adaptive accurate and fast vertical crowd counting via joint using depth and color information. *Multimedia tools and applications*, 73(1):273–289.

Ma, Z. and Chan, A. B. (2013). Crossing the line: Crowd counting by integer programming with local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2539–2546.

Nummiaro, K., Koller-Meier, E., and Van Gool, L. (2002). Object tracking with an adaptive color-based particle filter. In *Joint Pattern Recognition Symposium*, pages 353–360. Springer.

Van Oosterhout, T., Bakkes, S., Kröse, B. J., et al. (2011). Head detection in stereo data for people counting and segmentation. In *VISAPP*, pages 620–625.

Zhang, X., Yan, J., Feng, S., Lei, Z., Yi, D., and Li, S. Z. (2012). Water filling: Unsupervised people counting via vertical kinect sensor. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 215–220. IEEE.