International Journal of Advanced Smart Convergence Vol.7 No.3 8-14 (2018)
http://dx.doi.org/10.7236/IJASC.2018.7.3.8

IJASC 18-3-2

Development of People Counting Algorithm using Stereo Camera on NVIDIA Jetson TX2

Gyucheol Lee[1], Jisang Yoo[1], Soonchul Kwon[2]†

[1]Department of Electronic Engineering, Kwangwoon University, Seoul, Korea
[2†]Graduation School of Smart Convergence, Kwangwoon University, Seoul, Korea
gyucheol0116@gmail.com, jsyoo@kw.ac.kr, ksc0226@kw.ac.kr

Abstract

In the field of surveillance cameras, it is possible to increase the people detection accuracy by using depth information indicating the distance between the camera and the object. In general, depth information is obtained by calculating the parallax information of the stereo camera. However, this method is difficult to operate in real time in the embedded environment due to the large amount of computation. Jetson TX2, released by NVIDIA in March 2017, is a high-performance embedded board with a GPU that enables parallel processing using the GPU. In this paper, a stereo camera is installed in Jetson TX2 to acquire depth information in real time, and we proposed a people counting method using acquired depth information. Experimental results show that the proposed method had a counting accuracy of 98.6% and operating in real time.

Keywords: *People Counting, Stereo Camera, NVIDIA Jetson TX2, View Projection, Kalman Tracker*

1. Introduction

Surveillance cameras play an important role in the development of industry and security in modern society such as crime prevention, arresting criminals, collecting big data for marketing purpose. Recently, the intelligent surveillance camera industry has attracted much attention in response to product intelligence, which is one of the keywords of the 4th industrial revolution. An intelligent surveillance camera is a surveillance camera that includes an automation technology such as a surveillance camera that develops from simply shooting an image to determine whether a person himself / herself is suspicious of a person or an action or how many persons are present in a specific area.

People counting technology is a technique for analyzing camera images to determine the number of people in a specific area. As the history of surveillance cameras is old, there countless people counting techniques. However, most of the technology is based on a mono camera, and the level of the algorithm is simply designed due to the limitation of the hardware performance of the camera platform. Therefore, the detection

Manuscript Received: June. 18, 2018 / Revised: June. 25, 2018 / Accepted: June. 30, 2018
Corresponding Author: ksc0226@kw.ac.kr
Tel: +82-2-940-8637, Fax: +82-50-4174-3258
Graduation School of Smart Convergence, Kwangwoon University, Seoul, Korea

rate is reduced in a situation where a person is blocked by a person, and the accuracy of the counting is lowered. A technique developed to solve the problem of obscuration is to acquire depth information using a stereo camera. Distance information can improve the performance compared to the method using a mono camera by estimating the spatial structure around the camera. However, this technology could not be practically driven due to limitations of the hardware performance of the existing camera platform.

Recently, a high-performance embedded board with a GPU has been introduced, and these products have made it possible to substantially utilize algorithms with high computational complexity. Among them, the Jetson TX2 [1], launched by NVIDIA in 2017, was developed to run AI algorithms on embedded boards. This product has a CPU with a total of six cores, and a GPU with 256 cores is installed, enabling the parallel processing of algorithms using CUDA.

Therefore, this paper proposes a technique for counting people using a stereo camera in NVIDIA Jetson TX2. First, a stereo camera environment is configured using two cameras. Distortion of the stereo camera is corrected through camera calibration. We acquire the depth image through the stereo matching technique and project the side-view onto the top-view using depth information [2]. In the projected image, a person is detected using a height value and a projection ratio, and the detected object is tracked using a Kalman filter-based tracker [3]. Tracked objects are counted when they pass a line you specify.

The rest of this paper is as follows. Section 2 describes the proposed method. In Section 3, we show the experimental result. Finally, Section 4 concludes the paper.

2. Proposed Method

When installing CCTV, it is divided into side-view and top-view depending on the angle of the camera. The side-view is the angle of the camera diagonally, and the top-view is the vertical direction from top to bottom. The characteristics of the depth image are different according to each camera composition.

Since the depth image of the top-view is the closest distance between the head and the camera, the pixel corresponding to the head has the lowest value. This feature facilitates human detection, and a product has been released that actually counts people using a depth image in the top-view. On the other hand, the side-view differs from the top-view in that the depth value varies according to the distance between the camera and the object, so that the person cannot be detected only by the simple characteristic that the human head has the lowest value.

The proposed technique detects and tracks people in side-view and performs counting. A stereo camera is constructed to acquire depth images through a stereo matching algorithm. Use the depth information to project the side-view onto the top-view. This method solves the problem of undetecting due to occlusion in the side-view.

2.1 Stereo camera configuration

Two cameras are configured as a stereo camera as shown in Fig. 1. When installing the camera, the two cameras should be installed in parallel using the rig. In the stereo correction step, although the two images are horizontally aligned in software, when the horizontal state of the first two cameras is physically deviated greatly, an error occurs in the stereo correction step. Therefore, both cameras should be installed as horizontal as possible.

The distance between the cameras is set according to the maximum distance to recognize the pedestrian. Generally, as the distance between the cameras increases, the parallax of the left and right images of the object becomes larger, so that the parallax of the objects of a longer distance can be extracted well. In this paper, it is confirmed that the depth image is extracted up to 10m from the HD resolution when the distance between the cameras is 20cm, and the distance is set as the distance between the cameras.

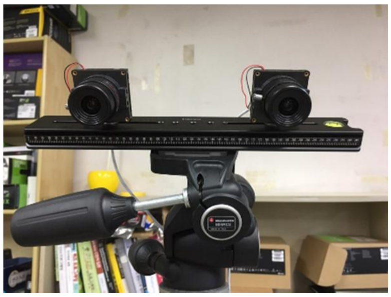


Figure 1. The stereo camera used in the proposed method

2.2 Camera calibration

The image acquired from the camera is affected by the camera internal parameters and distortion occurs [4]. The camera internal parameters are the focal length and the principal point. In order to perform image processing, distortion must be removed by using camera internal parameters. In order to acquire a depth image, stereo matching must be performed. In order to obtain a good depth image, the horizontal axis of the left image and the right image must be the same [5]. To align the two images horizontally, the epipolar line must be software-matched using camera external parameters. The camera external parameter is the relationship between the camera coordinates and the world coordinates as a rotation and movement matrix. Therefore, camera parameters must be acquired and this process is called calibration. The calibration proceeds through Equation 1, which describes the pinhole camera model [11].

$$s\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r1 & r2 & r3 & t1 \\ r4 & r5 & r6 & t2 \\ r7 & r8 & r9 & t3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{1}$$

where f_x and f_y denote the camera focal distances with respect to the x-axis and y-axis, respectively, and u_0 and v_0 denote the principal points of the camera with respect to the x-axis and the y-axis, respectively. r and t denote the elements of the rotation matrix and the movement matrix, respectively, and are external parameters of the camera. X, Y and Z denote points in the world coordinates, and x and y denote points in the image coordinates. s denotes weight of the image coordinate.

According to Eq. 1, knowing the pair of the coordinates of the image and the corresponding points of the world coordinate can acquire the camera internal parameters. To obtain the position of the point for each coordinate system, we use a uniform chess board. The corner point of the chess board is easy to extract because its characteristics are clear. The coordinates of the corner points extracted from the image are image coordinates, and the distance between the actual corner points of the chessboard is the world coordinate. By entering these two pairs of coordinates into the SolvePnP function in OpenCV [6], we can obtain camera internal parameters.

2.3 Stereo calibration and input of the counting line

Stereo matching is a method of acquiring parallax images by calculating distances of corresponding points in left and right images. Parallax images allow the acquisition of depth images using simple mathematical expressions [5].

In stereo matching, if the external factors except the camera-to-camera distances are unified, that is, if the two images are completely horizontal, finding the corresponding point becomes easier to find because it changes into one dimension. However, even if two cameras are installed in parallel, there is a slight physical difference. Therefore, it is necessary to make two images parallel in software. OpenCV [6]'s undistortStereoImages function makes it easy to get stereo-corrected images.

After completing the stereo correction, enter the counting line. Counts when a person crosses the counting line. A mono camera is counted when counting lines pass through the counting line because the counting line has only image coordinates, so the counting accuracy is lowered. On the other hand, the proposed technique uses the depth information to know the world coordinates for the counting line. In other words, counting is more accurate because you can determine if a person is passing through the counting line in space.

2.4 Disparity extraction by stereo matching

Extract disparity through stereo matching. Disparity means the distance of the corresponding points in the left and right images, and the unit is the pixel. Stereo matching technique uses various semi- global matching (SGM) [7] which is widely used in the proposed technique. SGM achieves disparity of high accuracy and algorithm structure is suitable for parallel processing, and GPU can be used for real time processing. Fig. 2 shows the disparity image extracted using SGM.

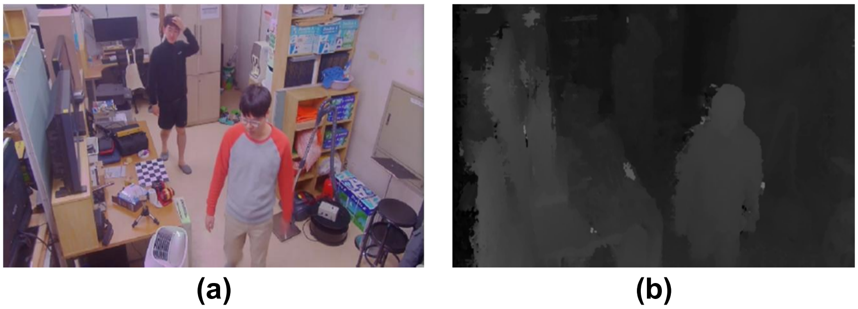


Figure 2. SGM result (a) Color image (b) Disparity image

2.5 Background removal

Since the top-view projection we apply to the side-view has a large amount of computation, if it is applied to the entire region of the depth image, the speed becomes very slow and it becomes difficult to perform in real time. Therefore, the top-view projection is applied only to the moving area, thereby greatly reducing the amount of computation.

We apply the background removal technique to the depth image to obtain the moving area [8]. The depth image has no shadow, but when the depth image is played back as a moving image, the object may be extracted as a moving region due to the shaking of the object contour. Also, in stereo matching, an error occurs in the disparity value in the region where there is no characteristic of the color image. When this region is reproduced as a moving image, an uneven depth value is outputted and it is detected as a moving region when the background removal technique is applied. These regions calculate the size and filter out the regions above the threshold.

2.6 View projection

The side-view is projected onto the top-view using the acquired depth and camera parameters. In order to project to the top-view, the rotation and movement matrix of the camera relative to the ground must be calculated. Before each of the matrices is calculated, the coordinates of the image corresponding to the actual ground and the corresponding world coordinate must be calculated.

If we look at the characteristics of the side-view, it contains a lot of ground because it is the view point in the diagonal direction. In particular, when the image is divided in half in the horizontal direction, the bottom region further includes a ground plane than the top region. Therefore, before the start of the program, the ground is entered as a square in the bottom area, and the center point of the rectangle is set as the origin of the world coordinate system. Then, the camera coordinates are converted with respect to the origin.

Fifty points excluding the origin are extracted at random from the ground, and camera coordinates for these points are acquired. Since the origin and 50 points are on the same ground, the Z axis value in the world coordinate system is zero. The X-axis and Y-axis values of the world coordinates for 50 points are expressed as the distance between the camera coordinates of the origin and the camera coordinates of the 50 points.

The image coordinate value and world coordinate value for 50 points are input to the solvePnP function of OpenCV to obtain the camera rotation matrix and movement matrix for the ground. The reason for extracting 50 points randomly is that a wrong matrix value can be obtained when referring to the depth value of the error in the depth image. We can obtain a matrix with high reliability by referring to a large number of points, but we could obtain good results by referring to about 50 through experiments.

After obtaining the camera rotation matrix and the movement matrix, the world coordinate value is obtained for the foreground obtained in the background removal process.

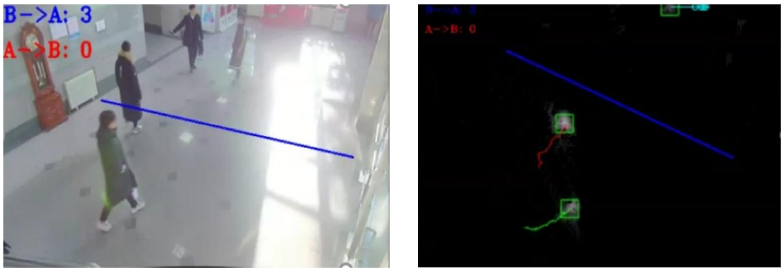
The world coordinate value of the foreground is relative to the previously set origin, and the X axis and Y value of the world coordinate system respectively generate a height map in which the horizontal, vertical and Z axis values of the image are composed of pixel values. The Z-axis value refers to the height of the ground since the origin is set to the ground.

Because height map alone cannot detect a person, an occupancy map is generated that represents how many pixels are projected in the height map coordinates.

We generate a likelihood map by modeling it as a Gaussian distribution to generate a map that can detect people by reflecting the height map and the occupancy map. When creating a Likelihood map, set the average height of the person to 160 and the standard deviation to 40, and set the objects with the corresponding height and deviation in the likelihood map to be detected with a high probability.

2.7 Object tracking and counting

Since the tracking algorithm using Kalman filter [3] uses only coordinate information without using the pixel information of the object, it has an advantage that the flight calculation amount is small in other tracking algorithm. Therefore, this problem is applied to the tracking algorithm using Kalman filter because it is aimed at driving in Jetson TX2. In addition, in side-view, the side-view is projected onto the top-view and then the object is tracked in the top-view. Therefore, a method of tracking position information using a Kalman filter is preferable. Fig. 3 shows the track results.



(a) **(b)**

Figure 3. Tracking result

3. Experimental Result

In this paper, we measure the accuracy of the counting of people using the proposed technique and the mono camera method. Because there is no official DB about the people counting, I directly shot the DB and measured the GT. When measuring a GT, if a person goes back and forth on the counting line, the program and GT are set to count only once beyond the first counting line since the counts will increase rapidly as the counting continues to update. Experiments were carried out at four resolutions (QVGA, VGA, HD, FHD) and camera parameters were changed when resolution was changed.

The accuracy of the counting accuracy was measured by counting GT of the whole image. Table 1 shows the counting accuracy at HD resolution. Because Axis-Cognimatics [9] uses a mono camera, accuracy is reduced when occlusion occurs. On the other hand, the proposed method shows higher accuracy than mono cameras because it solves the problem of masking by top-view conversion of the side-view.

Table 1. Comparison of counting accuracy

Method	Counting Accuracy (%)
Axis-Cognimatics	87.58
Proposed method	98.60

Table 2 shows the result of measuring the speed. In QVGA and VGA resolution, only CPU was used. In HD and FHD resolution, GPU was used in SGM process. Experimental results show that the speed up to HD resolution is more than 10 FPS which can be called in real time in CCTV field.

Table 2. FPS measurement result of proposed method for each resolution

Resolution	FPS (frame per second)
QVGA	30.3
VGA	12.4
HD	11.6
FHD	5.7

4. Conclusion

In this paper, we propose a technique for counting people using a stereo camera in NVIDIA Jetson TX2. We constructed a stereo camera environment using two cameras and projected the top-view to the side-view using the depth image acquired through the Semi-Global Matching. In the projected view, the person was detected through the height value and projection ratio. The detected persons were tracked using a tracker based on a Kalman filter [3] and set to be counted when a person crossed a user-specified counting line. Experimental results show that the performance of the proposed method is higher than that of mono camera.

References

[1] NVIDIA Jetson TX2 Module. *https://developer.nvidia.com/embedded/buy/jetson*-tx2

[2] T. Darrell, D. Demirdjian, Neal. Checka, and P. Felzenszwalb, "Plan-view Trajectory Estimation with Dense Stereo Background Models," *Proceedings of the International, Conference on Computer Vision*, Vol. 2, July 2001. DOI: 10.1109/ICCV.2001.937685

[3] Brown, R. Grover, and P. YC Hwang, *Introduction to Random Signals and Applied Kalman Filtering*, New York: Wiley, pp. 400, 1992.

[4] J. Weng, P. Cohen, and M. Herniou, "Camera Calibration with Distortion Models and Accuracy Evaluation," *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 14, No. 10, pp. 965-980, Oct 1992. DOI: 10.1109/34.159901

[5] A. Fusiello, E. Trucco, and A. Verri, "A Compact Algorithm for Rectification of Stereo Pairs," *Machine Vision and Applications*, Vol. 12, No. 1, pp. 16-22, Mar 2000. DOI: https://doi.org/10.1007/s001380050120

[6] G. Bradski, and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*, O'Reilly Media, pp. 555, 2008.

[7] H. Hirschmuller, "Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, July 2005. DOI: 10.1109/CVPR.2005.56

[8] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-Camera Multi-Person Tracking for EasyLiving," *Proceedings Third IEEE International Workshop on Visual Surveillance*, July 2000. DOI: 10.1109/VS.2000.856852

[9] Axis People Counter. https://www.axis.com/ko-kr/products/axis-people-counter

[10] Y.H. Hong, S.J. Song, and J. Rho, “Real-time Tracking and Identification for Multi-Camera Surveillance System,” *International Journal of Internet, Broadcasting and Communication(IJIBC)*, Vol.10, No.1, pp. 16-22, 2018.

[11] Martins, H. A., J. R. Birk, and R. Bo Kelley. "Camera models based on data from two calibration planes." Computer Graphics and Image Processing 17.2 (1981): 173-180.