Top-view People Counting in Public Transportation using Kinect

Filip Malawski
AGH University of Science and Technology
Faculty of Computer Science, Electronics and Telecommunication
Department of Computer Science
al. Mickiewicza 30, 30-059 Krakow, Poland
e-mail: fmal@agh.edu.pl

This article describes a method for people counting in public transportation. In this particular scenario, various body poses corresponding to holding handrails must be accounted for. Kinect sensor mounted vertically has been employed to acquire a database of images of 1-5 persons, with and without body poses of holding a handrail. An algorithm has been devised for robust people counting, consisting of multiple steps. The handrails are removed by substituting an average image of the handrails from the image with persons holding a handrail. The image is then processed in blocks in order to find potential local maxima, which are subsequently verified to find head candidates. Finally, non-head objects are filtered out, based on the ratio of pixels with similar and near-zero value, in the neighbourhood of the maxima. The method has an average accuracy of 91% and has proved to handle well the handrails in the depth maps.

Key words: people counting, Kinect, public transportation

Introduction

The problem of people counting is an important aspect of video surveillance. Possible applications include, among others, security on mass events, analysis of pedestrian traffic, tourist flow estimation, analysis of marketing efficiency in shops or malls. One particular application, that is the focus of this paper, is people counting in public transportation. Information on how many people travel in public transportation, depending on different routes and hours, is crucial for proper allocation of buses, trams and subways. In the big cities, with hundreds of thousands, or even millions of residents, and hundreds or thousands public transportation vehicles, proper management of the transportation network poses a difficult challenge. Obtaining information on traffic patterns by manual counting is expensive, time-consuming and inconvenient. As an alternative, automatic people counting devices could be installed in the vehicles and propagate the data in a real-time manner. This could in fact introduce new applications, such as providing the city residents with the information regarding current fill of the public transportation vehicles, so they could better plan their rides.

People counting is a well-known problem in the computer vision area and multiple solutions have been proposed so far. RGB, side-view images have been employed in several methods, such as blob descriptors [1] and fusion of shape and movement information [2]. However, they are dependent on illumination changes, occlusions and crowded environments. On the other hand, top-view depth images proved to be much more robust to these conditions. In [3] depth images are used for a coarse person segmentation and the RGB images are used for subsequent people tracking. In [4] authors compare the performance of a top-view people tracking system based on a Time-of-Flight (ToF) sensor with a system based on stereo camera. In [5] a method is proposed for people counting in top-view depth images by employing a water filling method. A new feature descriptor is devised in [6] for detection of human heads. In [7] and [8] methods based on fusion of depth and color information are explored.

Although some of the presented methods have very high efficiency (over 99%), they do not consider a scenario, in which people may be holding handrails, which is a typical situation in the public transportation. It has a considerable impact on the depth map and therefore on the efficiency of the people counting algorithms. In this paper a novel method is presented, based on top-view depth-camera images, which is robust to different arrangements of people in public transportation. The method employs local maxima detection with subsequent filtering of non-head objects, resulting in over 90% accuracy.

Methods

The study was conducted in two stages. First a database of top-view depth images with varying number of people in different body poses has been acquired. Secondly a method has been devised for people counting and verified on the database. The method is specifically tailored for the public transportation scenario in order to handle the handrails in the images. For the purpose of comparison, separate results

are provided for the method with and without the final filtering of non-head objects.

Database

For the acquisition of the depth images, Kinect camera has been employed. It has a 43^0 (vertical) x 57^0 (horizontal) field of view and provides 640x480 pixels depth images, with the rate of 30 frames per second. The device is equipped with an infrared (IR) emitter and sensor. The emitter projects a pattern of structured light which is then analyzed by the sensor in order to create the depth map.

The camera has been installed for top-view image acquisition, approximately 2,60m from the ground, and the images have been taken about 80-100cm from the subjects heads. There are 1-5 persons in each image and in half of the images there is at least one person with body pose of holding a handrail. Additional data has been obtained be rotating the recorded images. The database contains total of 500 images - 100 images per each number of people. Data acquired from the Kinect sensor has been calibrated to record images in such a range that only head, shoulders and arms of the subjects are visible. Everything that is further than that is not relevant in subsequent processing. A sample image from the database, containing 5 persons, is presented in fig. 1.

Head candidates detection

Proposed method is based on a parameter of expected size of human head. Given a setup with Kinect sensor installed on a known height it is possible to measure a minimal expected area that will be occupied by a single head. Once a local maximum is found it can be assumed that no part of other head may be present in the neighboring area of the selected range. In this study, based on series of experiments, the range of the maximum neighborhood was set to 60 pixels, which implies that minimum distance between two maxima was 120 pixels.

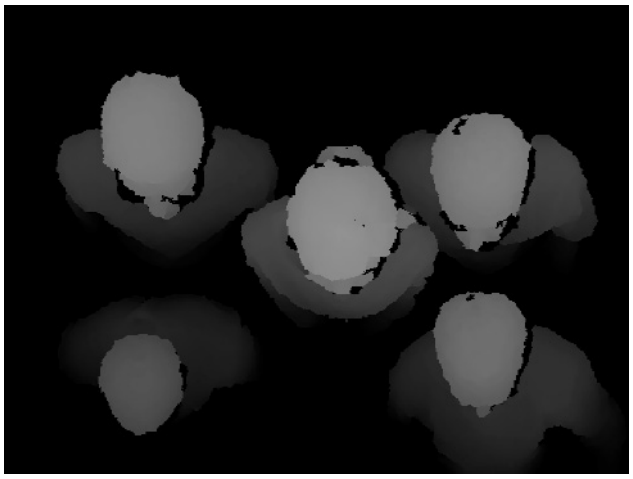

Figure 1. Sample top-view depth image of 5 people

Head candidates detection consists of 3 steps: 1) local maxima candidates are found in a block-processing manner 2) local maxima are selected which are potential head candidates 3) head candidates are fine-tuned.

In the first step the image is divided into blocks with size equal to the range of the neighborhood. Then, in each block, pixels with highest values (closest to the sensor) are found. Blocks, which contain only pixels with zero value, are omitted, as they are irrelevant for the subsequent processing. The selected size of the blocks guarantee that all potential maxima will be found. The grid of blocks as well as block maxima are presented in figure 2 (left). In the next step the potential maxima are verified. The area around the potential maxima, with the range of the minimal allowed distance, is scanned in order to check if this is in fact the pixel with the highest value in their neighborhood. This discards incorrect maxima and leaves only head candidates, as seen in figure 2.

Finally, the neighborhood is scanned in order to find other pixels with the same value as the maximum and an average position of all of these pixels is taken as the final head candidate. This allows to locate the head more accurately in case when the sensor, due to its limited depth resolution, sees the top of the head as an area of a few pixels with the same value.

Once a maximum is verified as a head candidate (see fig. 2 right), the area around it is marked as occupied by a head, therefore other maxima areas may not intersect with this one.

Filtering of non-head objects

So far we have not discussed the influence of the handrails on the depth images. Sample depth map containing a handrail is presented in figure 3 (left). Although it occupies a substantial part of the image, it can be easily removed (this step is performed before finding the head candidates). Once the sensor is mounted in the public transportation vehicle it takes a series of images of the space below it, with no people present. Therefore only handrails are visible, whose positions don't change. By substituting the image of the handrail (see fig. 3 middle) from any image with a per-

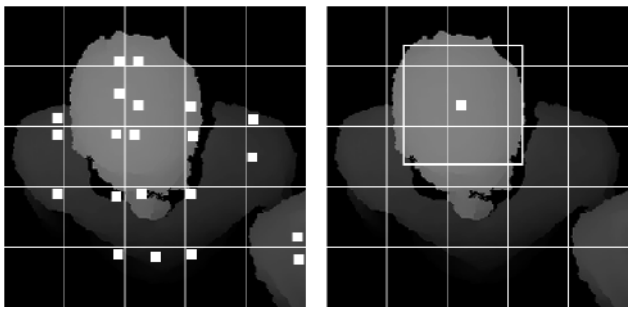

Figure 2. Part of a depth image divided into blocks, with block maxima marked (left) and head candidate selected (right)

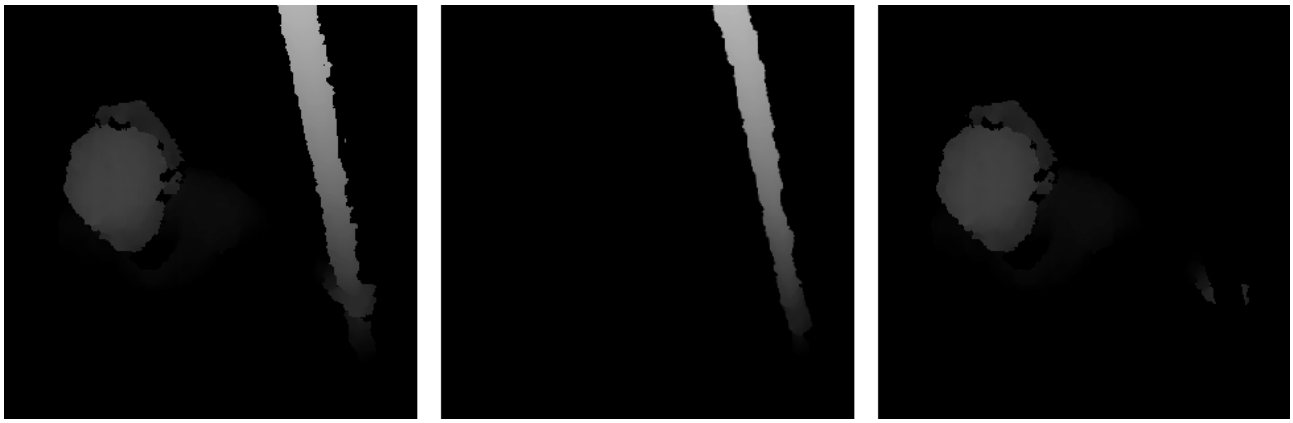

Figure 3. Handrail removal: image with person holding a handrail (left), image of the handrail (middle), first image with the handrail removed

son holding this handrail (see fig. 3 left) we can obtain an image with the handrail removed (see fig. 3 right). Due to the noise in the images of the handrail, it can occupy slightly different area in each depth map. The sensor has to take multiple images and then create an average, so the whole handrail will always be removed from subsequent depth maps.

Although the handrail is removed, the hand itself remains and would be recognized as another head candidate (see fig. 3 right). Therefore after the head candidates detection step, filtering is applied, in order to remove non-head objects such as stretched arms. Heads are verified by scanning the area around the maximum and computing the ratio of proper pixels to improper pixels. Proper pixels are the ones with value similar to the maximum, while improper ones have distinctly different value - mostly zero value, indicating a pixel out of the range of interest. The ratio distinguishes the heads well, as maxima corresponding to stretched arms and hands have usually much more improper pixels around them. The ratio threshold providing best classification of head and non-head maxima has been selected based on a subset of the dataset by iterative search with adaptive step.

Results

Proposed algorithm has been verified on the acquired database. Table 1. presents the results grouped according to the different number of people in the image. Two cases have been compared – in the first one the final filtering step is omitted and all the head candidates are counted. In the second case, non-head maxima are removed. Significant difference in the results can be observed for both cases, indicating that in this scenario it is crucial to handle the non-head maxima, as they can have great influence on the efficiency of people counting.

Without the filtering step the average result is close to 50%, which is due to the fact that half of the images in the dataset contain a body pose of holding a handrail. With the filtering the average accuracy is much higher - 91%. Although the handrails are handled rather well, there is still room for improvement for the head detection. In both cases – with and without filtering – the accuracy drops significantly, when there are five persons present in the image. As verified by visual inspection of the points marked as heads by the algorithm, these errors occur due to the shoulders being incorrectly recognized as heads.

Table 1. Accuracy of the proposed method, comparison of cases with and without final non-head filtering

# people in the image	accuracy [%] without filtering	accuracy [%] with filtering
1	50	100
2	59	99
3	47	89
4	49	94
5	32	73
average	**47**	**91**

Conclusions

In this paper a robust method of people counting in public transportation have been presented. A database of 1-5 persons, including images with people holding handrails, has been acquired and employed to verify the method. Average accuracy has been achieved of over 90%. The method has been proved to handle well the problem of different body poses, such as stretched arms, which are typical in this scenario. Additional work is required to fine-tune the algorithm to handle incorrect detection of the shoulders.

The method is based on block maxima detection and subsequent verification of head candidates. It works in real-time on low-end processors (1 GHz). Although the method is currently employed for static images only, it can be extended to process each frame from depth sensor and provide tracking of the people movements.

Acknowledgements

This research was partially supported by AGH-UST grant no. 11.11.230.124.

Bibliography

1. Yoshinaga S., Shimada A., & Taniguchi R. I.: „Real-time people counting using blob descriptor.„ Procedia-Social and Behavioral Sciences, 2(1), pp. 143-152, 2010
2. Patzold M., Evangelio R. H., & Sikora T.: „Counting people in crowded environments by fusion of shape and motion information.„ In Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on (pp. 157-164), Aug. 2010
3. Bevilacqua A., Di Stefano L., & Azzari P. „People Tracking Using a Time-of-Flight Depth Sensor." In AVSS (Vol. 6, p. 89), Nov. 2006
4. Bondi E., Seidenari L., Bimbo A. D., Bagdanov A. D.: „Real-time people counting from depth imagery of crowded environments." IEEE Computer Vision, 2003
5. Zhang X., Yan J., Feng S., Lei Z., Yi D., & Li S. Z.: „Water filling: Unsupervised people counting via vertical kinect sensor.„ In Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on (pp. 215-220). Sept. 2012
6. Rauter M.: „Reliable human detection and tracking in top-view depth images.„ In Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on (pp. 529-534), June 2013
7. Dan B. K., Kim Y. S., Jung J. Y., & Ko S. J.: „Robust people counting system based on sensor fusion.„ Consumer Electronics, IEEE Transactions on, 58(3), pp. 1013-1021, 2012
8. Wateosot C., & Suvonvorn N.: „Top-view Based People Counting Using Mixture of Depth and Color Information.„ The Second Asian Conference on Information Systems, ACIS 2013, October 31 – November 2 , 2013, Phuket, Thailand

MSc. Eng. Filip Malawski *– is a PhD candidate at AGH University of Science and Technology in the Department of Computer Science. He received his MSc degree in computer science at AGH University in 2012. His research interests include computer vision, image processing and human-computer interaction.*