Detection of people from time-of-flight depth images using a cell-tracking methodology.

Basavarajaiah S. Totada
Department of Electrical and Computer Engineering
The University of Texas at El Paso
El Paso, TX USA
bstotada@miners.utep.edu

Sergio D. Cabrera
Department of Electrical and Computer Engineering
The University of Texas at El Paso
El Paso, TX USA
sergioc@utep.edu

***Abstract*—This paper describes a method of detection and tracking of people using depth images captured by a Time-of-Flight (ToF) camera, such as those obtained with a Microsoft Kinect V2. Key advantages of this approach are that the identity of people is not disclosed, and the system can operate in low-light conditions. The automated approach developed here is inspired by cell-tracking methods as implemented in the well-known biomedical imaging software called CellProfiler. Our approach involves significant preprocessing of the depth images by a combination of adaptive center weighted median filtering and iterative inpainting. The next step is detection of each person's head using depth local minima information. The classification of each person is typically possible using evidence of shoulder depth information assisted by Laplacian of Gaussian (LoG) based matched filtering. After some additional processing and blob analysis, further quantification and monitoring of people is done using a multi-object tracking system based on Kalman filtering. The main applications are in the general area of collection of statistical information in smart building entrances/exits for security and commercial use.**

Keywords- Kinect depth sensor, occupancy detection, people counting, people tracking, time-of-flight, top-view camera.

I. INTRODUCTION

Body surveillance cameras in public areas play a very important role in public safety and security. More recent applications of cameras include their use to quantify and track pedestrian traffic flow. This application can be focused on the domain of *smart buildings* for counting people and measuring pedestrian traffic in and out of a building or between different sections of a building. However, very few studies focus on the use of a more recent type of camera, Time-of-Flight (ToF) cameras, which directly produce depth/distance images and videos where a person's identity cannot be determined.

This research has the following initial objectives in seeking to quantify individuals in a specific location or passing through an entrance or exit using depth images and video from a ToF camera:

- Accurate detection of people entering and leaving an area by monitoring their presence with an overhead camera.
- Detection of individuals with or without head accessories.
- Discrimination among individuals and objects even when they are close to each other.
- Differentiating a moving and a non-moving individual.

This research has many commercial significances in addition to surveillance, crowd control and disaster management. Our focus is on accurate counting and monitoring of individuals in a complex dynamic environment typical of a building entrance or exit. The identity of the individuals is not revealed due to the use of depth images rather than standard images. This secures the privacy of the individuals in the crowd.

Images and videos from overhead cameras are available for downloading from existing databases that include ToF camera sources, such as those used by the authors of [1]. Various approaches for preprocessing of depth images have been evaluated since these are known to be noisy and of low resolution, see the sample image in Fig.1. In this paper, state-of-the-art methods, which have been developed for conventional images and video, are applied to the ToF images and video from the GOTPD1 database [2] provided by the authors of [1]. These methods include Iterative Inpainting [3] and Adaptive Center Weighted Median Filtering [4]. We have adapted and improved general image and video processing methods to perform head detection on depth images using existing and tailored segmentation algorithms. For this, we observed how heads appear on depth images to identify key features that can be extracted for their detection. Motion between frames in a sequence of depth images is also exploited since people will be the moving objects in the scene. We also prototyped a solution using the CellProfiler software [5] which has been found to conveniently process images of a very similar nature to the ToF depth images of people from the top-view, including: Region of Interest (ROI) selection; Laplacian of a Gaussian (LoG) filtering producing cell-like objects from people in the scene; cell-like object detection and labeling; object tracking, etc. Using the tracking information, an algorithm has been developed completely in MATLAB to accurately identify and track new individuals entering the area under observation and similarly to identify the individuals leaving the scene.

II. ALGORITHM

A block diagram which shows each stage of the proposed algorithm is shown in Fig.2. This diagram is described next in brief but more detail is provided later in the paper.

A. Preprocessing

1) Frame extraction

Working with all the images in a video sequence cannot be developed and tested as a complete process. Thus, the first step we follow in our algorithm is to extract the video frames and then we can directly and independently edit the extracted images. The frames extracted from the existing database that includes ToF camera images [1], [2] are available in unsigned

integer 32 bit format which is then converted to unsigned integer 8 bit format, for the ease of processing in MATLAB.

2) Background removal

During surveillance of any area, the background is neutral

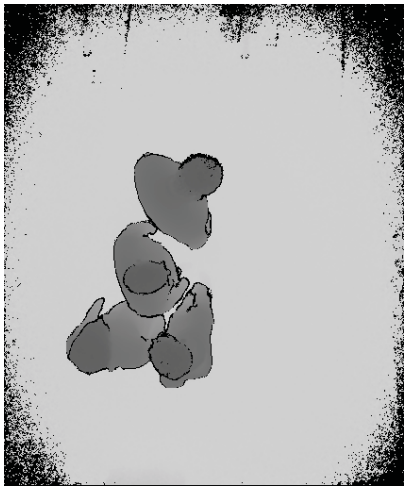


Fig. 1. Example of noisy depth image from Time-of-Flight (ToF) camera.

but subjects are not, hence image comparison would be the effective and easy method to remove the background. This can be done by comparing two images, one with the subjects and one without the subjects. Background removal can be done by replacing all the pixels which are nearly the same on both images with white pixels whose graylevel values are 255 and replacing the other pixels with those of the image with the subjects. The image comparison (or background removal) operation is defined as follows:

$$h(i,j)=\begin{cases}255, & \text{if} \quad f(i,j)=f_1(i,j)\\ f_1(i,j), & \text{Otherwise}\end{cases}$$

where, $h(i,j)$ is output image after image comparison, $f(i,j)$ is the depth image without the subjects, and $f_1(i,j)$ is the depth image with the subjects.

B. Noise reduction

Elimination of noise in the image is recommended as a prior step before subsequent processing (edge detection, image segmentation and object recognition) due to the nature of ToF images. Noise models relevant to this application can be the *fixed value* impulsive noise model (salt-and-pepper) or the *random value* impulsive noise (uniformly distributed) model. In this paper, the denoising of corrupted images is done by iterative inpainting and adaptive center weighted median filtering.

1) Systematic wiggling error

The main error source of a ToF sensor is systematic wiggling error. It alters the measured distance by shifting the distance information significantly towards or away from the sensor [6].

As explained in [7], ToF cameras work on the principle of pulse modulation. Distance from the light reflecting object to the ToF camera is calculated by measuring the phase difference between the reflected infrared light and reference infrared light. That is the reason why most cameras are equipped with active illumination units. The phase of the reflected signal is calculated as

$$\phi=\arctan\left(\frac{A_1-A_3}{A_2-A_4}\right),$$

where, A_1, A_2, A_3, A_4 are samples of signals shifted by 90°, and the distance d is proportional to phase ϕ which is calculated using signal frequency f_{mod} and speed of light c by $d=c\phi/4\pi f_{\text{mod}}$.

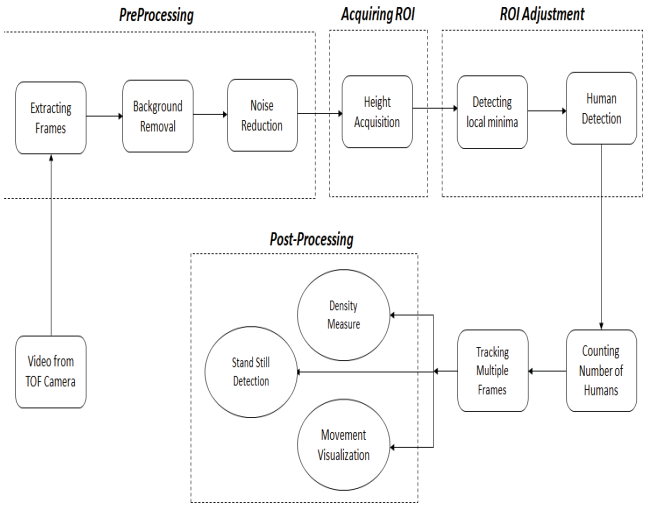


Fig. 2. Block diagram of the proposed TOF image processing system for people detection and tracking.

In the process of distance calculation from the phase difference, a systematic wiggling error arises. Due to hardware and cost restrictions, the theoretical assumption of sinusoidal signal shape is generally not suitable in reality [8]. As a result, the systematic wiggling error appears.

By visual analysis of the depth images, see the example shown in Fig.1, we can notice that clusters of pixels on the image boundary show large amount of systematic wiggling error. To eliminate this error, we define an ellipse from the center of the depth image whose perimeter touches the boundary of the depth image. We consider only the region within the ellipse as the effective area of view that we will use for depth images from the ToF camera.

2) Iterative inpainting

General applications of *inpainting* are image restoration, object removal, text removal, and special effects. Here we use a very simple inpainting approach to target the fixed value noise. First, we smooth the image using a gaussian filter which is a separable product of two one-dimensional gaussian functions, as shown in the equations below. Closely following reference [3], the mathematical description is as follows:

$$g(x,y)=\frac{1}{C}e^{-\left(x^2+y^2\right)/2\sigma^2},$$

$$H(x,y)=f(x,y)*g(x,y),$$

or,

$$H(x,y)=\sum_{j=1}^{height}\sum_{i=1}^{width}f(i,j)g(x-i,y-j)$$

where $H(x,y)$ is the output smoothed image, $f(x,y)$ is the noisy input depth image, and $g(x,y)$ is the two-dimensional gaussian filter impulse response normalized by C to produce a gain of 1.0 at zero-zero frequency.

Next, we replace the noisy pixels of the original image by the pixels from the smoothed version, as follows

$$f(i,j)=\begin{cases} H(i,j) & \text{if } f(i,j) \text{ is noisy} \\ f(i,j), & \text{Otherwise} \end{cases}$$

We have formed an iterative algorithm and each update gets us closer to the desired result awhere no change occurs after the replacement step. A simple threshold is used to declare which pixels are noisy since they are those with graylevel near zero (below 50) in these ToF images.

3) Adaptive Center Weighted Median Filter(ACWMF)

As described in [4], the adaptive center weighted median filter forms estimates based on the differences between the current/present pixel and the outputs of center-weighted median (CWM) filters with various center weights. It employs the switching scheme based on the impulse detection mechanisms, which are pixels that are considered outliers and should be changed by the filtering. It utilizes the center-weighted median filter with different center weights, which realizes the impulse detection by means of using the difference between the outputs of CWM filters and the present pixel being processed. The final output is switched between the median and the present pixel itself.

The ACWMF method includes a parameter s which according to the authors [4], good results can be obtained using $0 \le s \le 0.6$ for elimination of both types of impulse noises [9]. In this paper we use the parameters s=0.1, $[\delta_0, \delta_1, \delta_2, \delta_3]$=[40,25,10,5], and a neighborhood size of 3 by 3 in adaptive center weighted median filter to remove ToF image noise assumed to be modeled as random-valued impulse noise, see Fig. 3.

C. Human detection

1) Height acquisition

Throughout this paper we use the existing dataset acquired with a Kinect V2 ToF camera provided to us by the authors of [1]. This data is captured with a ToF camera located at a height of 3.4 meters in an overhead position, and its optical axis is perpendicular to the floor plane. As explained by the authors in [1], the coordinates from the camera are denoted as X_c, Y_c, Z_c, and its origin as O_c, see Fig. 4. By looking at the experimental setup, we know that the floor plane and camera plane are parallel, and the distance between them is the height of the camera from the floor which is h_c = 3.4 m. Let us consider a point p in the scene whose coordinates are X_p, Y_p, Z_p and let d_p be the distance between the point p and the center of the camera lens, which is the origin O_c. Let h_p be the height of the considered point p from the floor, then

$$d_p = \sqrt{X_p^2 + Y_p^2 + Z_p^2} \quad \text{, and} \quad h_p = h_c - Z_p \ .$$

So, it is possible to measure the height h_p of each point p in the scene from the acquired ToF image.

In this paper we define a height-of-interest that produces a region-of-interest (ROI) based on measured depths. The ROI corresponds to everything which is 2 feet (60.96cm) or higher above the ground and discard other depths. As an example, Fig. 5 shows the same image of Fig. 3 with only relevant gray levels corresponding to the ROI. All other gray levels are set to 255 for maximum contrast.

2) CellProfiler

We prototyped a solution using the CellProfiler software [5], which is an open source tool for conveniently processing and quantifying data from biological images. The motivation for using CellProfiler for human head detection is that there is only one head for each person as there is usually only one nucleus per cell.

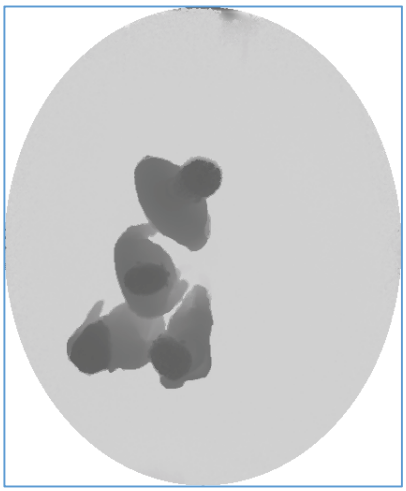


Fig. 3. Example of depth image after noise reduction .

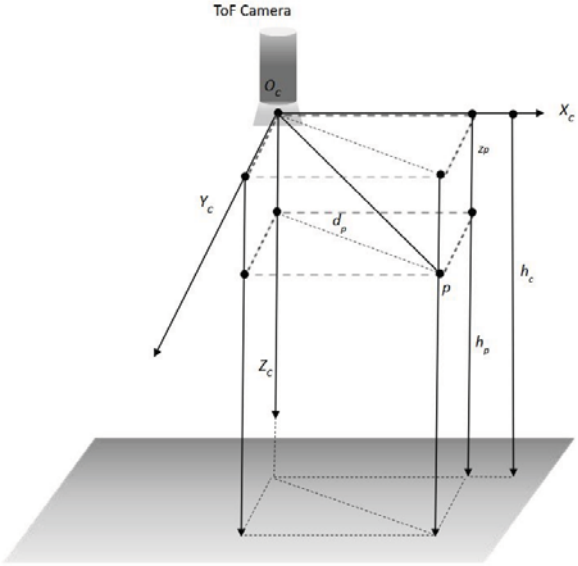


Fig. 4. Experimental setup defined as in [1].

Fig. 5. Depth image after height acquisition and keeping only the gray levels in a limited range of depths below a threshold (a desired ROI) which are the pixels closest to the camera.

The software contains already developed methods and advanced algorithms of image analysis which are user friendly. Thus, it is a flexible platform for testing and developing new methods of image analysis for images with cell-like objects.

In this software, image evaluation methods can be executed by pointing and clicking using CellProfiler's graphical user interface. The CellProfiler makes use of the concept of a 'pipeline' of modules. Each module processes the images in some manner, and each module is placed one below the other to create a pipeline. The pipeline used in our work is shown below in Fig. 6.

Over 50 CellProfiler modules are available at present and most of them are automatic, however, the software also allows interactive modules. Modules are mixed and matched for specific projects and we can alter performance of each modules by adjusting parameters appropriately in the module settings. Upon starting analysis, each image travels through the pipeline and is processed by each module from top to bottom, in order.

a) Errors of CellProfiler

After the process of height acquisition of all subjects above 2 feet from ground, if we input those images into CellProfiler, it may give out wrong results as shown in Fig. 7a). Here, even though CellProfiler identifies four different objects, it assumes two objects to be the same and gives them the same color code and label. Due to this, there are errors in the measurement of trajectories of objects which in-turn produce wrong tracking results. This error can be overcome by changing the height acquisition parameters.

If we acquire the region belonging only to human heads, which is done by height acquisition of all subjects above 4 feet and below 7 feet from ground level, and then input those images to CellProfiler, it now gives the correct measurements as shown in Fig. 7b).

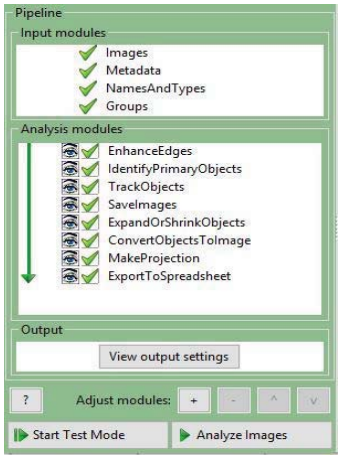
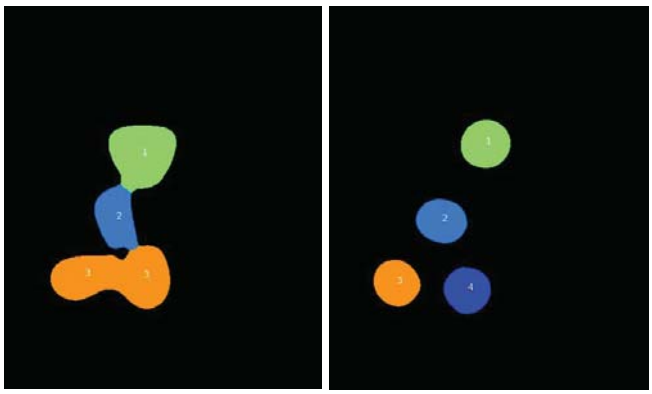


Fig. 6. A multi-step CellProfiler pipeline used in our work.

By looking at Fig. 7a), and Fig. 7b), we can conclude that CellProfiler works well only for the standing humans and those standing and with complements (hats, caps, backpacks, etc.). However, it cannot detect or track kids or adult heads that are lower in height, such as when people are sitting on wheel chairs. To overcome these limitations and to obtain more control over the processing, we move to a MATLAB implementation. The operations performed using MATLAB are described in the next sections of this paper.

3) Region of interest adjustment

In this project, the region of interest is every subject which is 2 feet and above from ground after background removal and noise reduction. Also, the region of interest adjustment in our case can be defined as elimination of all the other subjects except human because our intention is to detect, count, and track only humans. The procedure followed for region of interest adjustment is as follows. The human can be detected and differentiated from other subjects by looking at physical appearance of a human. Hence, we use the head and shoulders of a human body as the parts of interest to form a solution. This was motivated by the work in [1], and is done by following the steps explained below.



a) b)

Fig. 7. Output of TrackObjects module of CellProfiler a) with (left) and b) without (right) error.

a) Step1: Detection of local minima

The position of the camera (overhead) is such that the nearest part of a human body to the camera is the head. Therefore, the heads of humans will automatically have the lowest gray level values since they have the smallest distance (depth) to the ToF depth camera. Hence, by finding the local minima of the acquired depth image we can detect the human heads, this is done in MATLAB software by using the inbuilt function "imregionalmin".

b) Step2: Searching interest height.

This step increases the robustness of the algorithm by verifying whether each local minimum identified in the previous step corresponds to human head or not.

To identify the subject present in a depth image as a human, the criterion is that the height difference between the shoulder and the head should be in a prescribed range h_i. In our work, based on anthropometric considerations [11], we have selected interest height (height difference between shoulder and head) to be in the range of 30cm to 40cm, and we have set the typical area covered by a human head as equal to 15×15 cm.

We know that, due to overhead position of the camera all human heads are detected as local minima, but all the local minima identified do not correspond to human heads. To verify whether the identified local minima correspond to human heads, we use the above set values. Also, in order to

search for shoulders, we define 8 different radial directions δ_i ($1 \le i \le 8$) around each local minima as shown in Fig.8.

Initially, we find the height h_m of each identified local minima and then search for interest heights h_i around that local minima in the predefined 8 different radial directions δ_i at a distance of $l/2$. If we find the interest height (shoulder) h_i in any of the predefined directions δ_i, we can conclude that local minima corresponds to a human head.

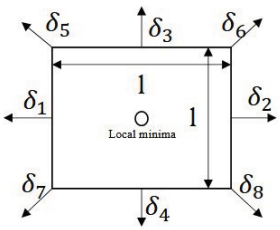


Fig. 8. Predefined 8 radial directions around local minima.

Additionally, to achieve a correct segmentation for a reliable classification process, it should work even if people are close to each other or if there are partial occlusions. To achieve this in our work, if the region with interest height h_i is associated with one local minimum (head) then that region's measurements will not be considered for any other local minimum. This will guarantee us at least one-pixel distance between identified subjects.

Only the area covered by heads of size $l \times l$, and the identified shoulders in particular directions δ_i of local minima, which have been determined to correspond to human heads, will be retained in the region of interest and all other subjects will be discarded. Fig. 9, shows the depth image after region of interest adjustment.

D. *Counting and tracking humans*

Laplacian of Gaussian (LoG) filters are second derivative filters preceded by gaussian filtering. Hence, an LoG filter is useful for finding edges (regions of rapid change in gray level) in an image. It can also be used in matched filtering to find similar shapes in an image, as done in CellProfiler. The derivation of the LoG impulse response is well known [10] and the normalization used by Matlab involves the same constant C used in the gaussian filter, see above, thus

$$LoG(x,y) = \frac{1}{C}\left[\frac{x^2+y^2-2\sigma^2}{\sigma^4}\right] e^{-\left(x^2+y^2\right)/2\sigma^2} \quad .$$

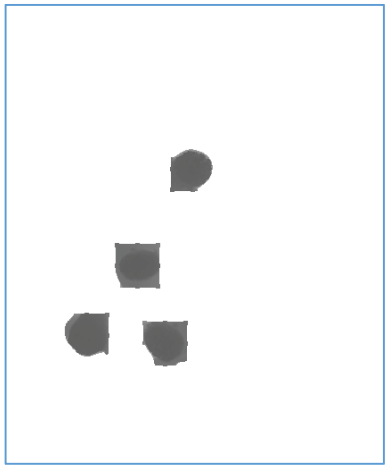


Fig. 9. Example of depth image after region of interest adjustment.

In our work, we use the inbuilt MATLAB function "fspecial" to produce a discrete impulse response for the LoG filtering. After detecting the edges, we use the inbuilt MATLAB function "imfill(image,'holes')" to fill holes present in the image. The output image after this operation will be a binary image. This operation is performed because later it will be easy to get measurements of identified objects. Fig. 10, shows the identified objects.

Next, we use the inbuilt function in MATLAB "bwboundaries (image, 'noholes')" to count the number of objects present, which is nothing but the number of humans. And then, we apply inbuilt function in MATLAB "regionprops (label_matrix, 'centroid')" and "centroid = stats (object_numb).centroid" to find the centroids of all objects identified in each frame (which are the positions of humans in each frame), and save them in the cell. So, we can track a person by plotting the centroids of each frame and assigning a z-axis value equal to the frame number. Fig. 11 shows the tracking of a single regular person moving around in a scene. Also, we can do tracking by loading images (after object detection) into the TrackObjects module of CellProfiler.

Fig. 10. Example of objects identified in depth image.

III. RESULTS

The robustness of the proposed algorithm is demonstrated by running different sequences of data through it, see [12] for more detail.

A. *People detection accuracy*

Fig.12a) is the output obtained when an image sequence of a single regular person moving randomly in a scene was given as input to the proposed algorithm. Fig. 12b) is the output obtained when data sequence of eight persons moving very close to each other in a scene was given as input to the proposed algorithm.

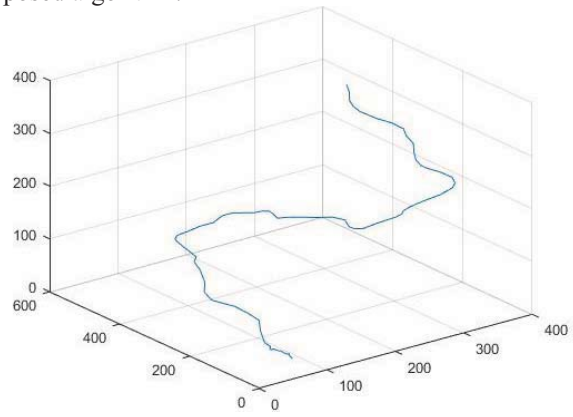
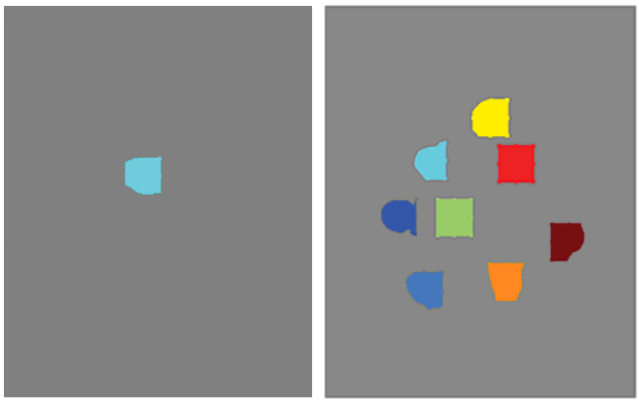


Fig. 11. Example of trajectory of a subject identified in a depth image.

The people detection accuracy of the proposed algorithm is defined as the ratio of subjects detected to subjects present. The accuracy of detecting a single person in a scene is equal to 97.50% and the accuracy of detecting multiple people in a scene is equal to 98.12% .



a) b)

Fig. 12. Detection of a) single person present in the scene and b) multiple people present in the scene.

B. *Tracking*

We have tracked persons by plotting the stored values of the centroids of all objects identified in each frame and assigning the frame number to the z axis. Fig.13 shows the comparison of the track obtained from our algorithm versus the track obtained from ground truth value of the dataset for a single regular person moving randomly in a scene.

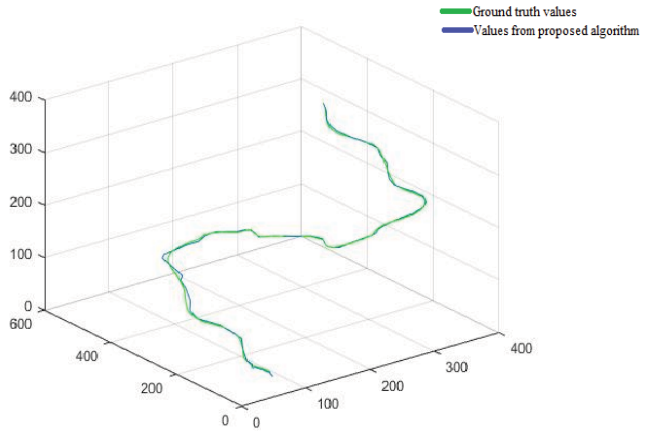


Fig. 13. Comparision of track obtained from our algorithm versus track obtained from ground truth values of the trajectory available in [2].

The root mean square error (RMSE) of our algorithm is found using the coordinates of the ground truth locations and the centroids obtained by our algorithm. The result for the tracking of Fig. 13 is a RMSE value of 4.2027 pixels.

IV. Conclusions

In this work, we proposed a new method for the robust detection of people in depth images, captured by an overhead ToF camera. The CellProfiler image processing software was initially used to prototype an approach using a methodology of cell-tracking. Our algorithm was then completely coded in MATLAB. The proposal has several stages, and it allows detection of multiple people even with components like hats, caps, bag packs, etc. Firstly, we remove background and define Region of Interest (ROI). Then, we find local minima and adjust a Region of Interest by considering human upper body geometry. Finally, using inbuilt functions of MATLAB we count the number of people present in the scene. The proposal also includes saving centroids of detected people, from which we can have information of trajectory and velocity of each person. Additional related work can be found in [13]-[15].

Acknowledgment

We would like to acknowledge support from the Texas Instruments Foundation Endowed Scholarship Program for graduate students in electrical and computer engineering at the University of Texas at El Paso. We sincerely thank the authors of [1] for providing us with the GOTPD1 dataset.

References

[1] Luna, Carlos A., C. Losada-Gutierrez, D. Fuentes-Jimenez, A. Fernandez-Rincon, M. Mazo, and J. Macias-Guarasa. "Robust people detection using depth information from an overhead Time-of-Flight camera." *Expert Systems with Applications* 71 (2017): 240-256.

[2] Macias-Guarasa, Javier, Cristina Losada-Gutierrez, David Fuentes-Jimenez, Raquel Garcia-Jimenez, Carlos A. Luna, Alvaro Fernandez-Rincon, and Manuel Mazo. "GEINTRA overhead ToF people detection (GOTPD1) database description." (2016), available at this link http://www.geintra-uah.org/archivos/GOTPD1/GOTPD1-readme.pdf .

[3] Richard, Manuel M. Oliveira Brian Bowen, and McKenna Yu-Sung Chang. "Fast digital image inpainting." In *Appeared in the Proceedings of the International Conference on Visualization, Imaging, and Image Processing (VIIP 2001), Marbella, Spain*, pp. 106-107. 2001.

[4] Chen, Tao, and Hong Ren Wu. "Adaptive impulse detection using center-weighted median filters." *IEEE Signal Processing Letters* 8, no. 1 (2001): 1-3.

[5] Carpenter, Anne E., Thouis R. Jones, Michael R. Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A. Guertin et al. "CellProfiler: image analysis software for identifying and quantifying cell phenotypes." *Genome biology* 7, no. 10 (2006): R100.

[6] Lindner, Marvin, Ingo Schiller, Andreas Kolb, and Reinhard Koch. "Time-of-flight sensor calibration for accurate range sensing." *Computer Vision and Image Understanding* 114, no. 12 (2010): 1318-1328.

[7] Hagebeuker, Dipl-Ing Bianca, and Product Marketing. "A 3D time of flight camera for object detection." *PMD Technologies GmbH, Siegen* (2007).

[8] Kweon, In-So, Jiyoung Jung, and Joon Young Lee. "Noise aware depth denoising for a time-of-flight camera." In *20th Korea-Japan Joint Workshop on Frontiers of Computer Vision*. Asian Federation of Computer Vision (AFCV), 2014.

[9] Chen, Tao, and Hong Ren Wu. "Adaptive impulse detection using center-weighted median filters." *IEEE Signal Processing Letters* 8, no.1 (2001): 1-3.

[10] Gonzalez, Rafael C., and Richard E. Woods. Digital image processing third edition." *Pearson Prentice-Hall,* (2008).

[11] Chen, Hsiao-Lin, and Dengchuan Cai. "Body dimension measurements for pillow design for Taiwanese." *Work* 41, no. Supplement 1 (2012): 1288-1295.

[12] Totada, Basavarajaiah S., People Detection from Time-of-Flight Imagery with Inpainting-based Preprocessing, master's thesis supervised by S. D. Cabrera, Dept. of Electrical and Computer Engineering, University of Texas at El Paso, August 2018.

[13] Del Pizzo, Luca, Pasquale Foggia, Antonio Greco, Gennaro Percannella, and Mario Vento. "Counting people by RGB or depth overhead cameras." *Pattern Recognition Letters* 81 (2016): 41-50.

[14] Owodolu, A. A., C. Bolu, A. A. Abioye, and K. U. Efemwenkiekie. "Development of a System for Counting of People Using MultiCamera and Sensors." In *IOP Conference Series: Materials Science and Engineering*, vol. 413, no. 1, p. 012001. IOP Publishing, 2018.

[15] Stahlschmidt, Carsten, Alexandros Gavriilidis, Jörg Velten, and Anton Kummert. "Applications for a people detection and tracking algorithm using a time-of-flight camera." *Multimedia Tools and Applications* 75, no. 17 (2016): 10769-10786.