3D Object Detection and Tracking in an Indoor Enviroment

K. Amplianitis, M. Adduci, and R. Reulke

Humboldt-Universität zu Berlin, Institut für Informatik, Computer Vision Group
Unter den Linden 6, 10099 Berlin
eMail: reulke@informatik.hu-berlin.de
URL: http://www.informatik.hu-berlin.de/cv

Abstract. Human extraction and tracking is an undergoing field were many researchers have been working on for more than 20 years. Although several approaches in the 2D domain have been introduced, 3D literature is limited, requiring further investigation. Within this framework, an accurate and fast-to-implement pipeline is introduced working in two main directions: pure 3D foreground extraction of moving people in the scene and interpretation of the human movement using an ellipsoid as a mathematical reference model. The proposed work is part of an industrial transportation research project whose aim is to monitor the behaviour of people and make a distinction between normal and abnormal behaviours in public train wagons using a network of low cost commodity sensors such as Microsor Kinect sensor.

1 Introduction

Human detection and tracking has been a challenging task for many scientists in the computer vision and machine learning communities. Many researchers have been thoroughly working in the direction of improving and refining existing algorithms for achieving minimum detection failures. To the best of our knowledge, the majority of these methods use training data for learning a classifier capable of detecting and also labelling a human posture or action. Extending the problem in 3D, the works of [9], [3], [11] and [5] involved detecting people and their body parts taking advantage of the richness of the RGBD data. Nevertheless, these approaches seem to deliver poor detection rates in environments with lots of noise in the cloud, fast illumination changes and overcrowding.
Interesting work was also introduced in the 3D people tracking literature: the Unscented Kalman Filter [13] and the Random Hypersurface Models [1] are some of the most recent development techniques applied in the area of human tracking in a cloud. In a multi Kinect sensor configuration, the work of [4] proposed a method for detecting and tracking a person in the scene by fitting a cylinder shape to its body.
For the specific application we are interested in, these approaches would fail for the following reasons:

- The Kinect network configuration in the wagon covers only a limited field of view (FOV), introducing large amount of noise in the depth images due to the conflict between the infrared emitters. Thus, the generated point clouds contain a lot of noise which in turn force the algorithms to fail even after extensive

tuning of the parameters.

- A train wagon consists of many non reflecting areas such as windows, dark color seating, etc. that significantly reduce the amount of data in the point cloud. Areas in which the infrared light of the sensor is absorbed by the element of the object, returns no data to the depth image.
- Instant illumination changes (e.g. entering a station platform from a dark tunnel) are some natural environmental conditions that strongly effect the quality of any detection algorithm.
- Rush hours in early morning and late afternoon introduces a lot of occlusions and overlapping between people in the wagon, making it impossible to detect any human instance.

In this proposed work, we try to address these issues by currently improving the work of [6] which is based on pure 3D background estimation between an empty background and a current processed cloud. From the extracted foreground, an ellipsoid is utilized for encapsulating each individual body. One main advantage of this mathematical shape (compare e.g. to a sphere) is the fact that an ellipsoid can better represent a human figure and can be used to derive larger amount of information from it (higher degrees of freedom).

2 Approach

We introduce an approach for extracting, monitoring and tracking human figures in 3D space coming from RGBD Kinect sensors. Main objective is to exploit a mathematical representation such of an ellipsoid for obtaining meaningful information from the human posture. At first, raw point clouds are acquired from all RGBD sensors through a synchronized camera acquisition system. For computational efficiency, every incoming point cloud is trimmed in the depth direction based on a predefined threshold. Subsequently, a background subtraction is performed using the octree approach of [6]. If moving objects are present, foreground points are projected on a 2D binary image and connected components is used for preserving contours with an area larger than a predefined threshold. The rest of the blobs are considered to be noise and therefore are removed.
The remaining parts of our algorithm entails the fitting of an ellipsoid over every human figure in the scene and approximately monitor his behaviour through the underlying geometry of the shape. The algorithmic part of the ellipsoid, for consistency and clarity is examined in a separate chapter whereas the rest of the steps are extensively analysed in the current section.

2.1 Point Cloud Trimming

It is unlikely that all points of a point cloud are required for extracting foreground moving objects. In most cases, points placed outside the region of interest can be removed so that the remaining part of the work flow could be accelerated. The trimming is been done in the depth direction, where is more likely to have points that are closer to a wall or any kind of object that does not contribute to the rest of the scene.

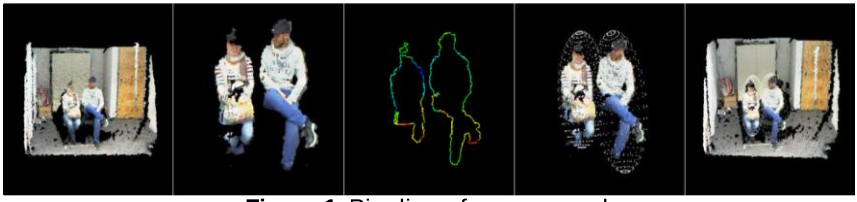

Figure 1. Pipeline of our approach

2.2 Background Subtraction

We use the method of [6] for extracting moving objects in the scene. It works by recursively encoding the structural differences between the octree representations of two point clouds. These structural differences represent the spatial changes between the two clouds which in our case is the moving foreground. An octree is a tree based data structure in which every internal/leaf node has exactly eight children. Each node in the octree subdivides the space it represents into eight octans. In the case of object extraction it can be used for detecting spatial changes between the octree of the background and current cloud. Spatial changes in the leaf node of the tree (sparsity of points, amount of neighbours, etc.) can give an indication of these spatial changes. Depending on the predefined size of the leaf node, detection sensitivity rate and processing time can vary. Large leaf nodes are faster to process but don't provide detailed information on the foreground and therefore only very significant spatial changes are detected. On the contrary, very small leaf sizes can capture detailed spatial changes but the computation time is extremely costly. In all cases, based on the FOV and amount of detection required, leaf size can be adjusted manually by the user. For more information refer to the author's paper ([6]).

2.3 Projection on a 2D plane

Extracted moving objects from the scene in a traditional background estimation fashion are always followed by some surrounding noise. Instant illumination changes and shadows are some the most common problems which still remain unsolved even in the 2D domain. In 3D space, depending on the cloud generation source (stereo cameras, TOF, structured light sensors), noise modelling differs. We approach the problem by projecting all 3D foreground points on a 2D binary image and extracting all contours using connected component analysis. Contours which have a size larger than a predefined threshold are retained and the rest are removed. Performing an accurate calibration of the sensor will definitely affect the quality of the projection. Therefore, a pre calibration step is strongly suggested in this case.

2.4 Convex hull of a human figure

In the field of computational geometry, convex hull of a shape is the smallest convex polygon containing all points of that object. Considering this statement, the

ellipsoid computation would only require the convex hull of the body rather than the complete set of points representing the body. If all points had to be used, computational speed would significantly drop, keeping the performance of the algorithm in very low levels. Mathematical notation of the convex hull and its use within this framework is given in the next chapter.

3 The ellipsoid as a human motion interpreter

As was stated in a previous section, an ellipsoid can better approximate the human shape compare than a sphere due to its shape and high degrees of freedom. Inspired by the work introduced in [8] and [12], we were able to fit a minimum enclosing ellipsoid to the extracted human figure and monitor his behaviour through the geometrical variations of the ellipsoid. The attributes that were tracked are the following:

- Center of gravity of the ellipsoid
- Vertex position
- Semi-major axes length
- Variation of each semi major axis expressed in percentage
- Rotations omega, phi and kappa extractred by the covariance matrix of the ellipsoid
- Volume of the ellipsoid
- Constrained angles omega, phi and kappa as dipicted by figure 2.

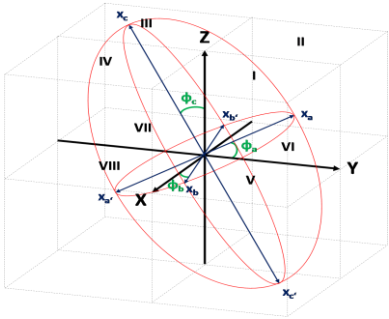

Figure 2. The constrained ellipsoid

We deploy a three dimensional Cartesian right-handed coordinate system in which any set of two lines are perpendicular to each other and have a length equal to one (refer to Fig. 2). The main idea is to have a coordinate system placed at the center of gravity of the current ellipsoid remaining invariant to ellipsoid variations. In this way, any rotation that the ellipsoid undergoes due to the human pose change, this coordinate system will continue to preserve a fixed orientation and therefore every semi-major axis of the ellipsoid will be checked against a predefined axis of this system.

The complete pipeline for retrieving the angles of each semi-major axis with respect to this "imaginary" fixed coordinate system is defined as follows: As an input to the algorithm, the position of the vertices is given. Next step involves finding the angle of every semi-major axis, translated and normalized at the origin with respect to a predefined axis of the fictitious system. As a final step, finding the octant area in which every normalized vertex falls into, some logical statements - constrains for the derived angles are made.

Assigning a reference coordinate axis of the fixed system to each semi-major axis of the ellipsoid was chosen based on what is considered as *human approximated zero angle movement*. Approximated zero movement is represented by a human posture when he's standing with his hands down. Therefore, for fixed axis *X* the

semi-major axis *b* is assigned, also characterizing the width of the person. Then, the *Y* axis is related to the *a* axis which corresponds to the depth of the person and finally the *Z* axis is referred to the *c* axis which expresses the height of the body. After that, a check is been done in to all three angles in order to ensure that they will always lie between the range of $-180 \leq \varphi_a, \varphi_b, \varphi_c \leq 180$. Regarding the octans orientation, every octant has its own placement in the coordinate frame depending from the sign of the reference axes. Therefore, *first octant(I)* is placed where x, y and z values are positive and *last octant(VIII)* where all points are negative. The rest of the octants are numbered based on a counter clockwise rotation around the positive z axis as seen in Figure 2.

4 Experimental Results

4.1 Camera configuration and hardware

A train wagon was provided as a prerequisite to the project by a transportation firm for acquiring, testing and evaluating different algorithms. The area of interest was surrounded by a network of four Kinect sensors, mounted on an aluminium construction as depicted in figure 3(a). Due to a non-disclosure agreement (NDA), we are currently not able to publish results from the wagon, therefore a simulated environment (replica) was build within a room using the same construction frame and similar texture/environment characteristics as the one of the wagon 3(b) covering a FOV of approximately 10 square meters. Scenarios, similar to the ones acquired in the wagon where also generated in the room, containing one to more people in normal or abnormal state. Acquisition was done in parallel by all sensors with an acquisition rate of approximately 19 fps. Every sensor is connected to a dedicated USB bus due to the high rate of information generated from both infrared and RGB camera.

(a) (b)

Figure 3. 3(a) Camera mounting configuration within the train wagon and 3(b) in the simulated environment.

One of the main drawbacks of using multiple structured light sensors is the drastic reduction of the depth image quality due to the intersection of near-infrared light in space. Therefore, all sensors were oriented towards the lower center of the scene restricting the overlapping only in the lower part of the FOV.

Concerning hardware performance, computers are configured with an Intel Core i7-3770 processor, 16GB RAM and a Samsung 840 Pro SSD. In present state, the complete framework is only able to run offline while real time processing would require better hardware performance but also further software optimization. Data from all sensors are processed with a frame rate of approximate 2 fps.

4.2 Calibration and bundle block adjustment

There are several libraries (eg. OpenNI, Freenect) which provide out-of-the-box calibration parameters of the Kinect sensor. Nevertheless, for achieving maximum possible accuracy of the generated point clouds, a more precise calibration is required. Main advantage of the Kinect sensor is that it uses low distortion lenses with faintly apparent displacement errors around the corners/edges of the images. The calibration was performed by using a regular chessboard with pattern size 2cm and inner dimensions of 5x7 rows and columns respectively. Since infrared and RGB sensors cannot work simultaneously, they were triggered to switch on and off continuously (a switch lasts 0.5 seconds), in order to acquire roughly the same chessboard data from the different perspectives. Detection and acquisition of chessboard points was done in a live mode using OpenCV's chessboard corner detector, which also delivers subpixel accuracy. The lenses were modelled using Brown's 10 parametric model [2]. To avoid any disturbances of the speckles coming from the infrared emitter in the infrared camera, the emitter was covered with tape and a external hydrogen lamp was used for detecting the chessboard corners. A total amount of 100 images was acquired and split (using a random selection algorithm) in 10 different sets of 24 images each, each of which the calibration was performed independently.

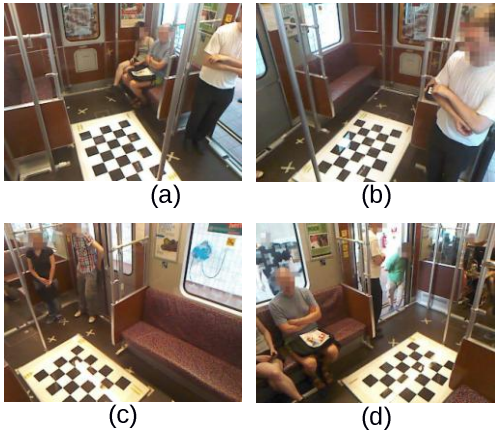

Figure 4. Regular chessboard used as a reference system for all sensors mounted in the train wagon.

Figure 5. From top row to bottom: raw scenes from different scenarios; foreground masks extracted by our approach; results from [6] approach; cloud to cloud background subtraction using a global threshold; The foregrounds extracted by our approach together with their encapsulated ellipsoids; ground truth masks generated by the implementation of [10] in the OpenNI framework.

Due to the multi-camera configuration in the wagon, every sensor produces different results which in turn can contribute for improving the quality of the extracted foreground (e.g. registration of all foregrounds, solving occlusion problems, etc.). Although current working status involves processing all sensors in parallel applying the introduced algorithmic pipeline to every sensor, results are automatically transformed into a common coordinate system for better comparing the data between them but also, in long term, fuse all information in a multi-sensor approach. The *Efficient Perspective n Point* algorithm [7] was used for transforming all sensors into a global coordinate system defined by a large chessboard with pattern size of 15cm (see Fig. 4). The accuracy of the camera external parameters (reprojection error) was in the range of less than a quarter of a pixel. Finally, we improved the accuracy of the camera's poses by setting the results from the previous step as approximate initial values to a photogrammetric bundle adjustment. Internal parameters of the sensors remained fix (due to accurate lens correction parameters) and only the external orientation parameters were refined delivering a variance of the unit weight, $\sigma_0 = 0.16$ pixels. Ground control points were generated setting the Z value to zero (coplanar reference object) and X-Y values according to the number of rows, columns and pattern size respectively. The main reason for using this form of rerefence system is the fact that it can be easily used as a reference object for all cameras. On the other hand, coplanar objects lack of spatial distribution information and introduce several geometrical constrains.

4.3 Object extraction and tracking

Different scenarios, similar to the ones in the train wagon, were captured in a simulated train field for testing and evaluating the quality performance of the algorithm intorduced in section 2. Our approach was checked against [6] and a common cloud to cloud background subtraction using a global distance threshold, setting the borderline between foreground and background. To the best of our knowledge, there are no other background subtraction approaches that could be compared against ours as most of them are heavily dependent on machine learning algorithms. Ground truth was generated by detecting the human figure from the depth images using the skeleton tracking algorithm implemented in the OpenNI framework [10]. The extracted figure was then projected into 3D space using the internal calibration information of the corresponding sensor. Results from different camera views and scenarios are given in Figure 5.
It is clear that our method outperforms the two other approaches, producing better quality foreground masks in all cases. All parameters were empirically defined after extensive evaluation and testing: for the octree, it was important to provide a leaf size that controls the amount of voxels in the cloud and was set to 0.10m. Then, the depth trimming of the point cloud was performed using a pass through filter preserving all points up to 4m. Also, contours on the binary image that had less than a 1000 or more then 7000 pixels respectively were removed. Finally, the global distance threshold for the cloud to cloud subtraction was set to 5cm.

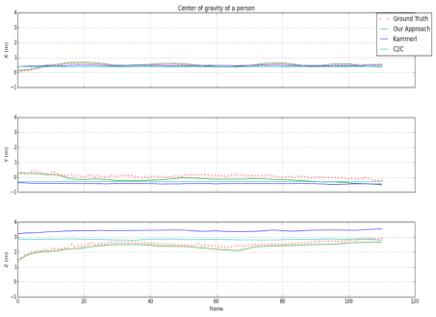

Figure 6. 3D trajectory of the center of gravity of a person as computed by the ellipsoid, projected in X,Y and Z planes.

Figure 7. Likelihood of the distance error for every point in the trajectory with respect to its ground truth.

Last step involves fitting an ellipsoid around the human figure and extracting its geometrical characteristics over time. Kalman filter was applied to all attributes of the ellipsoid for removing any unwanted sparks and smoothing out the data. Figure 6 shows 110 frames from a trajectory of a person as computed by the aforementioned approaches together with the ground truth generated from [10]. It is clear that our method produces greater stability compare to the other two methods as they tend to follow a constant plateau effect. This is because the amount of noise in the scene does not allow the ellipsoid to be encapsulated only around the body

but also incorporating the noise around it. On the contrary, our approach follows the ground truth trajectory in a more likewise manner. The trajectory of every approach was checked against the ground truth using the following formulation:

$$L^t = \frac{\left\| P^t - P_g^t \right\|_2}{\left\| P_g^t \right\|_2} \quad (1)$$

where L^t is the likelihood (in %) of every point on a trajectory at time t against its equivalent ground truth point, $||.||_2$ represents the second Euclidean norm and P_t, P_g^t are points on the trajectory of any of the two approaches and ground truth respectively. Figure 7 clearly shows the quality of likeliness between different approaches with respect to the ground truth. Our pipeline provides less than 20% likelihood fitting to the ground truth, where the rest tend to be far away from it, as a result of severe noise in the environment. This is observed in the areas higher than approximately 30-40% which means that the distance of a point in the trajectory is approximately a quarter away compare to the distance of the ground point from its natural zero origin.

One of the main drawbacks of our approach is the instant increase of the size of the ellipsoid when two or more people come very close to each other. Although this is controlled given a minimum and maximum size of a contour, it still remains an unsolved issue and it's currently investigated. All parameters of the ellipsoid are saved in an XML file and imported in a tracking visualizer for monitoring the behaviour of people in the train. Unfortunately, this visualizer was developed by another partner within the project and therefore due to NDA we are not yet allowed to make any results publicly available. Finally, psychologists in the social and cultural anthropology field where responsible for interpreting and classifying the behaviours as normal or abnormal.

5 Discussions and conclusions

This paper introduced a method of extracting, monitoring and tracking people in an indoor train environment using a network of sensors, were current state of the art machine learning detection approaches would fail due to the challenging environmental perturbations. Current state of the work involves processing all cameras in parallel using the algorithm presented in section 2. Results show that the proposed method can deliver high quality foreground segmentation masks compare to the ones of [6] and cloud to cloud subtraction. We were able to eliminate the noise and preserve only the moving person in the scene by modifying the approach of [6]. Results in the previous section also showed that the accuracy of the foreground strongly reflects on the accuracy of the ellipsoid. Noise in the surrounding can provide misleading information which does not help the monitoring process and eventually will result false interpretation of the behaviour. We where able to achieve a likelihood rate of less than 20% from the ground truth in comparison to the other approaches, most of the time retaining a deviation larger than 30-40% from ground truth. We also tried to filter out these noisy blobs from the processed clouds using different 3D filters but in all cases the resulting foreground

was very much affected by the noise in the scene.
In the preprocessing steps, calibration was mandatory for maximizing reliability of the produced results. The internal parameters were mainly used for generating the point clouds but also for projecting the 3D points on a binary image as discussed in section 2.3. Bundle adjustment was performed keeping the internal parameters fixed in the convergence process optimizing only the external values of the cameras.

References

1. Baum, M. and Hanebeck, U. D. (2013). *Extended object tracking with random hypersurface models*. CoRR.
2. Brown, D. C. (1971). *Close-range camera calibration*. Photogrammetric Engineering, 37(8):855–866.
3. Buys, K., Cagniart, C., Baksheev, A., De Laet, T., De Schutter, J., and Pantofaru, C. (2014). *An adaptable system for rgb-d based human body detection and pose estimation*. J. Vis. Comun. Image Represent.
4. Faion, F., Baum, M., and Hanebeck, U. D. (2012). *Tracking 3D Shapes in Noisy Point Clouds with Random Hypersurface Models*. In Proceedings of the 15th International Conference on Information Fusion (Fusion 2012), Singapore.
5. Hegger, F., Hochgeschwender, N., Kraetzschmar, G., and Ploeger, P. (2013). *People Detection in 3d Point Clouds Using Local Surface Normals*, volume 7500 of Lecture Notes in Computer Science, book section 15, pages 154–165. Springer Berlin Heidelberg.
6. Kammerl, J., Blodow, N., Rusu, R. B., Gedikli, S., Beetz, M., and Steinbach, E. (2012). *Real-time compression of point cloud streams*. In IEEE International Conference on Robotics and Automation (ICRA), Minnesota, USA.
7. Lepetit, V., F.Moreno-Noguer, and P.Fua (2009). *Epnp: An accurate o(n) solution to the pnp problem*. International Journal Computer Vision, 81(2).
8. Moshtagh, N. (2005). Minimum volume enclosing ellipsoids.
9. Munaro, M., Basso, F., and Menegatti, E. (2012). T*racking people within groups with rgb-d data*. In IROS,, pages 2101–2107. IEEE.
10. Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., and Blake, A. (2013). *Efficient human pose estimation from single depth images*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12):2821–2840.
11. Sigalas, M., Pateraki, M., Oikonomidis, I., and Trahanias, P. (2013). *Robust model-based 3d torso pose estimation in rgb-d sequences*. In The IEEE International Conference on Computer Vision (ICCV) Workshops.
12. Todd, M. J. and Yildirim, E. A. (2007). *On khachiyan's algorithm for the computation of minimumvolume enclosing ellipsoids*. Discrete Appl. Math., 155(13):1731–1744.
13. Ziegler, J., Nickel, K., and Stiefelhagen, R. (2006). *Tracking of the articulated upper body on multi-view stereo image sequences*. In CVPR (1), pages 774–781. IEEE Computer Society.