A High Reliability 3D Object Tracking Method for Robot Teaching Application

Yan-Chun Chen, Wei-Chan Weng, and Shi-Wei Lin
Metal Industries Research & Development Centre in Taiwan
E-mail: strongerlin@mail.mirdc.org.tw

Abstract. 3D Object tracking is the task of capturing the 3D position and pose of an object from each time-series image frame. As we known, 3D sensing technique can be realized by stereo vision, structured light, and ToF (time-of-flight) camera. All of them can capture the point cloud data for describing the depth information in a workspace. In past research, the reliability in 3D object tracking was a big problem for real industrial application, therefore, we address a different way to enhance the tracking accuracy and stabilize the tracing path for raising the reliability. In order to build a 3D tracking model and the workspace environment, we adopted an RGB-D camera which is the Intel® RealSense™ D400 Series depth modules to collect the cloud point data and RGB values. The built 3D tracking model should contains the information which includes points, normal and texture for producing many 2D object images with different perspectives. Then the produced images were fed to a SSD (single-shot detector) neural network to learn the object’s features for 2D tracking. In dynamic tracking process, the image frames were through the semantic image segmentation by DeepLabV3+ for only extracting the object information without hands and background. Thus, the reserved data only included object’s cloud point data and texture information in workspace. Then we use the iterative closest point (ICP) algorithm and the RGB intensity correlation method to confirm the object’s position and posture in workspace. The result shows that our method has a better performance than SSD method for tracking a self-predefined object.

1. Introduction
The 3D object pose estimation and tracking are widely implemented in computer vision and machine vision. Computer vision is processed nature problem, such as video surveillance system, virtual reality and augmented reality. The video surveillance system is designed to capture sequence frames using the camera, then we analysis the behavior of humans, car, and animals. Analytical approaches in 2D image were often adopted to track specific object technologies for notification or warning purposes in many papers. Medical care or terrorism and attack prevention is a popular application in this method, however, the object’s posture in 2D image is difficult to detect in a slight change. Therefore, 3D object pose estimation techniques are necessary to be developed for tracking and estimating objects’ pose in realistic street conditions [1]. 3D pose estimation was also applied on virtual reality and augmented reality which emphasize the human interaction in a virtual environment that can be extended to the applications of education, games and medical rehabilitation [2-5]. Machine vision is used for analysing product defects and guiding robotic arm to execute a pick & place mission in industrial applications. In the past, machine vision always delivered a single point data to robot arm’s controller to execute a movement program. Nowadays, the trend is that computer vision integrated
with robotic arm will make them friendlier for operators, and object tracking and pose estimating [6] will be the key.

Well image pre-processing can obtain good results for detecting an object’s trajectory and posture. Object detection method can be divided into four types [7], including point detectors, segmentation, background modelling and machine learning. The first type is point detectors which means interest point detectors, such as harris interest point detector [8], SIFT detector [9] and moment matrix computed with uniform Gaussian scale-space [10]. The second type is traditional classic numerical method for segmentation such as active contours [10], graph cut [11], etc. In this methods, image’s color-scale, texture, position by using statistical methods to achieve the purpose for segmentation. Recently, machine learning was also adopted to solve the segmentation problems [12]. In the third type, the background modelling was used for separating dynamic objects and static background. Generally, the target object which we need to detect will move continuously in natural environment, like human, animals, cars, and so on. Background modelling is often used for pedestrian detection in the surveillance field. But there is a disadvantages in this method, the object will be regarded as same as background when it moves very slowly. The classic background modelling methods includes mixture of gaussians model [13], dynamic texture background [14]. The last type is machine learning technique. Machine learning can be divided into supervised classifiers and unsupervised classifiers. Supervised classifiers must use the pre-marked data for analysing, such as support vector machine (SVM) [15] and Neural network [16-17]. Unsupervised classifiers analysis are based on the characteristics of the original data by using special statistical methods [18]. At present, machine learning can be applied on different field problems with the newest development of neural network models.

Our research dealt with the 3D tracking and pose estimating by RGBD image for a specific object which we predefined. The 3D rotation and translation can be obtained through depth information. We adopted two kind of techniques in this research, including feature-base and model-base[19]. Feature-base is used for tracking by the object’s texture and feature description through Statistical methods, such as kalman filter [20] and particle filter [21]. Model-base is used for considering the edge and geometry of object and also matches the features description. In object tracking, template matching [7] for considering the 6D pose estimation and the most famous method is the ICP( iterative closest point) method [22]. Although there are many estimation methods which was mentioned, the performances are still not good as ICP.

This paper proposes a reliable analysis procedure to solve 3D object tracking and pose estimating by using a single RGBD camera. In Section 2, we introduce different method for capturing the depth image and 3D image process method in this research. In Section 3, the result shows that our method has a better performance than SSD method for tracking a self-predefined object.

2. Methodology

Our approach is to process the xyz-rgb data from an RGBD camera. We remove background data and only retain dynamic data by a threshold value from the depth information. Then we will discuss how to create a complete object and region of interest (ROI) by image segmentation method. Finally, we use a 3D template model to match the dynamic image for tracking and pose estimating by ICP algorithm.

2.1. Background removal from depth information

In robotic vision, to select a right hardware can simplify many problems and get better results in different operation environments. Usually, an RGBD camera can catch the depth information for processing and calculating objects’ trajectory, position relationship, and 3D environment construct. Depth information can be obtained by different techniques for computing the distance on each pixel from cameras. The techniques include [24] Stereo vision, Time-of-Flight (ToF) and Structured light. Stereo vision is to calculate the disparity between the camera and image according to the edge information. Moreover, this system must be calibrated before using, the captured images should be consider the sufficient texture features and ambient light source. They always add a light source with some image pattern to enhance edge and surface information.
ToF means time of flight. The camera adopt active sensors that determine the depth value by measuring the time obtained by emitted IR light with high frequency to the object and back to the sensors. The depth information obtained by ToF is more accurate than Stereo vision. However, the noise generated during reflective process must be improved by conventional filters.

A structure light 3D image device composed of pattern projector, and a camera. The depth values are calculated from the reflective pattern imaged from objects and correlates it against the original pattern on a plane. The disadvantage of this technology has a short working distance, but it has the highest resolution of depth information than others.

In this paper, we must consider the resolution and accuracy of depth value, therefore the structure light based RGBD camera was used for obtain the xyzrgb data. In order to obtain the cloud point data of the dynamic object from the camera, the depth information extract out by (1).

\[ f(x, y) = \{ I_{RGB} | z \geq T_d \} \]  

In (1), where \( f(x, y) \) is foreground image. The foreground image was extracted by setting a threshold \( T_d \) was set for each position of the depth information \( z \).

2.2. Object segmentation

We used deep learning for image segmentation that must be considered the spatial resolution problem. These problems can be improved by pooling and stride. Moreover, the category classification error I is caused by object pose change. It can be solved by atrous convolution and Encoder-Decoder neutral network architectures. In traditional classification structure, the feature map of the last layer size is 1/32 than original image size after pooling. This size is not good for pixel by pixel classification results. The atrous convolution method can process larger range of resolution of an image, and it does not need to change the input and output channels by this kernel. This method adopt various convolution kernels with insert holes to create different feature maps for replacing the last layers of the network. Defined a 1D signal Equation as (2):

\[ y[i] = \sum_{l=1}^{L} x[i + r \cdot l] w[l] \]  

Assume that a one-dimensional signal is input \( x[i] \), dilated rate parameter \( r \), output \( y[i] \),and \( L \) of filter size. For instance, we use this convolution method on ResNet's network structure as shown below:

![Figure 1. Consider atrous convolution in ResNet to increase the reception field.](image)

In the original architecture, the last layer of the network is block4. The blocks 5 to 8 are copied from the block 4 to get more multi-scale cascade context (Fig1). Finally, this segmentation approach is used for dealing with multi-scale problems by the spatial pyramid pooling method with atrous convolution. That is shown below:
In Fig. 2, the Arous Spatial Pyramid Pooling (ASPP) is an architecture with different kernel size by various dilated rate. Besides, the architecture also includes an $1 \times 1$ convolution and an $1 \times 1$ pooling. The image was reconstructed by using two $4 \times$ deconvolution mask and added the low-level information.

2.3. Tracking and pose estimating
For tracking a dynamic object, we converted 2D object image to 3D point cloud. It was considered the camera's internal parameters after the image segmentation of our method. The internal parameters format as below (3):

\[
\mathbf{C} = \begin{bmatrix}
    f_x & 0 & c_x \\
    0 & f_y & c_y \\
    0 & 0 & 1
\end{bmatrix}
\]  

(3)

In (3), the $f_x$ and $f_y$ refer to the camera's x and y focal length. The $c_x$ and $c_y$ refers to camera's aperture.

\[
x = (u - c_x) \cdot \frac{z}{f_x}
\]

(4)

\[
y = (v - c_y) \cdot \frac{z}{f_y}
\]

(5)

\[
z = \frac{d}{s}
\]

(6)

In Eqs. (4), (5) and (6), the $x$, $y$, and $z$ are coordinates of 3D point cloud. The conversion calculation is through the coordinates of $u$, $v$, and $d$ on the image with camera internal parameters. The 3D point cloud is the coordinate set of space merge color information, such as RGB values. We improve processing efficiency by using three-dimensional coordinate in our approach. Suppose the two sets are $P$ and $Q$ that used euclidean distance calculate the most similar points in the two sets.

\[
d(\mathbf{p}_i, \mathbf{q}_i) = \left\| \mathbf{p}_i - \mathbf{q}_i \right\|
\]

(7)

Afterward, we consider translation and rotation changes for two sets of similar points and establish a translation matrix $T$ and a rotation matrix $R$. The two matrix use the least squares iterative variation to find a set of the most similar $R$ and $T$ matrix. According (7), it establish the objective function of the least square access as equation (8).

\[
E = \sum_{i=1}^{N} \left\| \left( R_{\mathbf{p}_i} + T \right) - \mathbf{q}_i \right\|^2
\]

(8)
3. Experiments

In this paper, the tracking method was implemented by using TensorFlow 1.09 and cuDNN 9 on python and the device is PC with Intel Core i5 CPU and NVIDIA Titan GPU card. Our primary aim is to achieve tracking the posture of hand tools in a video. The image with the size of 1280×720 pixels in the video analyzed can be divided into three categories including equipment-base, console-base, and operator-base.

The experiment is to compare the detection accuracy between the proposed method and the single shot multi-box detector (SSD). The training data from the video contains 600 frames including two types of hand tools. There are two different strategy, proposed method and SSD method for comparison in this paper:

1. Strategy by using SSD model:
   i. Use MobileNetV2 to separate features.
   ii. Train the SSD model.
   iii. We got the best result after iterating 2500 epoch.

2. Proposed strategy:
   i. Separate object into front and background part by using depth image, and remove background image—equipment.
   ii. Use PASCAL VOC 2012 database for training segmentation model to separate operator and hand tool in the front part of the image. After 2000 iterations we got the best result.
   iii. Finally, cloud points of the hand tool are left in the image only. We use ICP function in the MVTec’s software product HALCON to track the posture of a hand tool.

Following two figures are the results after SSD method and out method.

![Background1 (equipment-base)](image1)

![Background2 (console-base)](image2)

![Background3 (operator-base)](image3)

(a)Our Method  (b)SSD

Figure 3. The illustrations show that SSD and proposed methods to detect object in different backgrounds.
The Fig. 3 Shows the analyzed frames in different backgrounds which are equipment-base, console-base and operator-base. It shows that our method has robust results for tracking and pose estimating in different backgrounds. The score of detection rate in SSD and proposed method are listed in Table 1, and the scores prove that our method is better than the SSD method.

Table 1. Each space is (correctly framed /total frame number) in two method.

| Background Type        | Method          | Our Method     | SSD Method    |
|------------------------|-----------------|----------------|---------------|
| Background1 (equipment-based) | 0.81(300/369)  | 0.135(50/369) |               |
| Background 2 (console-based)  | 0.75(250/330)  | 0.181(60/330) |               |
| Background 3 (operator-based) | 0.92(210/226) | 0.442(100/226)|               |

As shown in Fig 4(a), 4(b), 4(c), our method performs well while changing to a metal tool or including multi-objects in a frame. In Figure 4(d) and 4(e), the hand tool’s posture and the moving path are shown by using HALCON’s ICP function.

4. Conclusion

The strategy combined both using 3D depth information and 2D object segmentation can improve the detection accuracy. Especially, our segmentation model without need retrain to achieve robust results in complex background or scale changes. Our method has a better performance than SSD while in the different environments, changing hand tools’ posture and multi-objects detection in a frame. However, this method cannot do well with highly reflective or slender objects. We will find another way to improve these problems in the future. For example, the high-reflection objects will be suppressed by image enhancement method.

5. References

[1] M. Andriluka, S. Roth, B. Schiele, Monocular 3D pose estimation and Tracking by Detection, CVPR(2010)

[2] T. Darrell, B. Moghaddam, A.P. Pentland, Active face tracking and pose estimation in an
interactive room, CVPR(1996)

[3] F. Ababsa, M. Mallem, Robust camera pose estimation using 2D fiducials tracking for real-time augmented reality systems, VRCAI(2004)

[4] I. Rabbi, S. Ullah, A Survey on augmented reality challenges and tracking. Acta Graphica, 24, 29-46(2013)

[5] K. Pauwels, L. Rubio, J.D’iaz, E. Ros, Real-time model-based rigid object pose estimation and tracking combining dense and sparse visual cues, CVPR (2013)

[6] T. Drummond, R Cipolla, Real-time visual tracking of complex structures, IEEE Transactions On Pattern Analysis And Machine Intelligence, 24,7(2002)

[7] A.Yilmaz, O. Javed, M. Shah, Object Tracking: A Survey. ACM Computing Surveys, 38(4),1-45 (2006)

[8] C. Harris, M. Stephens, A combined corner and edge detector, 147-151(Alvey Vision Conference, 1988)

[9] G. Lowe, Distinctive image features from scale-invariant keypoints. Int.J. Comput. 60(2), 91-110(2004)

[10] M. Kass, A. Witkin, D. Terzopoulos, Snakes: Active Contour Models, 1(4), 321-331(1988)

[11] I. Xu, N. Ahuja, Object contour tracking using graph cuts based active contours. ICIP. 277-280(2002)

[12] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deep lab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv:160600915(2016)

[13] C. Stauffer, W. Grimson, Learning patterns of activity using real time tracking. Transactions On Pattern Analysis And Machine Intelligence. 22(8),747–767(2000)

[14] A. Monnet, A. Mittal, N. Paragios, V. Ramesh. Background modeling and subtraction of dynamic scenes, ICCV, 13–16(2003)

[15] C. Papageorgiou, M. Oren, and T. Poggio, A general framework for object detection, ICCV(1998)

[16] A. Rowley, S. Baluja, T. Kanade. Neural network-based face detection. IEEE Trans. Patt. Analy. Mach. Intell. 20(1), 23-38(1998)

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, SSD: Single shot multibox detector. ECCV(2016)

[18] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis. IEEE Trans. Patt. Analy. Mach. Intell. 24(5), 603-619(2002)

[19] B. Seo, H. Park, J. Park, S. Hinterstoisser, and S. Ilic. Optimal local searching for fast and robust textureless 3d object tracking in highly cluttered backgrounds. IEEE Transactions on Visualization and Computer Graphics, 20(1), 99-110(2014)

[20] T.J. Broida, R. Chellappa, Estimation of object motion parameters from noisy images.IEEE PAMI, 90-99(1986).

[21] C. Choi, H. I. Christensen, Robust 3D visual tracking using particle filtering on the special euclide an group: A combined approach of keypoint and edge features. Int. J. Robot.Res., 31(4):498–519(2012)

[22] A. Tagliasacchi, M. Schroeder, A. Tkach, S. Bouaziz, M. Botsch, M. Pauly, Robust articulated-icp for real-time hand tracking, Technical report (2015)

[23] L. Maddalena, A. Petrosino. Background subtraction for moving object detection in rgbd data: A survey. Journal of Imaging, 4(5), 71 (2018)